universität wien

DISSERTATION

Titel der Dissertation

„Inference on a large number of hypotheses based on limited samples –
some points to consider“

Verfasserin

Magistra Alexandra Goll

Angestrebter akademischer Grad

Doktorin der Sozial- und Wirtschaftswissenschaften

(Dr. rer. soc. oec.)

Wien, im März 2008

Studienkennzahl lt. Studienblatt:	A 084 136
Dissertationsgebiet lt. Studienblatt:	Statistik
Betreuer:	o.Univ.Prof. Dr. Peter Bauer

Contents

List of Figures

List of Tables

1 Preface

In gene expression or proteomic studies mostly a large number of hypotheses are investigated. As compared to the large number of hypotheses (genes or proteins) to be tested, only for a small number of hypotheses noticeable effects exist. Two problems that may arise in this context are:

1. finding differentially expressed genes (proteins) among a large number of hypotheses and
2. finding prognostic scores to predict the clinical outcome of future patients.

This thesis consists of two parts. The first part investigates the problem to select the few genes (proteins) with an effect among up to thousands of candidates. Due to limited resources, the number of observations per hypotheses in conventional single-stage designs are low which limits the power. It has been shown that two-stage designs are a good option to improve the power. In these sequential designs, the first stage is used to screen for the promising hypotheses, which are further investigated in the second stage. Two-stage pilot designs only use second stage data and two-stage integrated designs use pooled data from both stages for the final test decisions on the screened hypotheses.
In genomic or proteomic studies there is an increasing focus on using a less accurate assay in early stages and a more accurate one in later stages. This thesis more thoroughly investigates this type of two-stage designs where the costs per measurement and effect sizes differ between the first and second stage. To compare different designs it is assumed that the total costs of the experiment are fixed. Both integrated and pilot designs are based on procedures either controlling the Family Wise Type I Error Rate (FWER) or the False Discovery Rate (FDR).

Asymptotically optimal designs will be derived and their statistical properties will be described.

Two scenarios are considered: In the first scenario the same method is applied in both stages but different costs per measurement may arise at both stages because specific experimental devices have to be produced at higher costs per measurement for the selected markers at the second stage. Furthermore the robustness of the optimal two-stage designs against misspecifications in the planning phase with regard to the proportion of true null hypotheses and the true effect size is investigated. In the second scenario the experimenter from the beginning may have the choice between two methods that differ in costs and effect sizes (a low-cost standard method or a high-cost improved method).
As extension, cases under less stringent distributional assumptions like unknown variance and correlation between hypotheses are investigated. Finally the constraint of integer stage wise sample sizes is discussed.

In the second part of the thesis, the problem of selecting and estimating a score from a large set of gene or protein measurements to allow prediction of a clinical outcome is discussed. Such a situation arises, e.g., if genetic measurements are available in samples of patients responding or not responding to a particular therapy respectively and these samples are used as a training set to construct a score for prognosis of the response of future independent patients. The prognostic ability of scores developed in this way will be studied for different selection methods. We tackle the question what we can expect, if in samples of patients responding and not responding to a particular therapy, markers are selected and used to construct a score for prediction of response in future patients. Prognostic scores based on three selection procedures are discussed:

1. Multiple testing of individual hypotheses controlling the FDR
2. Selecting the k "best" markers
3. Using a stepwise selection procedure

We simulated scenarios for different choices of the FDR and k and for different thresholds for a forward logistic regression. The predictive ability of such estimated scores is investigated

by simulating their Receiver Operating Characteristics (ROC) in an independent future patient. The different selection procedures are compared using the area under the ROC (AUC). We investigated the predictive abilities under the assumption of independent hypotheses with known variance. Additionally the situation is considered that cross validation is used to determine decision boundaries for the test based selection procedures optimal with regard to the area under the ROC-curve (AUC), thus achieving some information on the extent of false positive decisions.

Outline of the thesis:

The first part of the thesis is based on the following published paper:

> Goll and Bauer (2007): Two-stage designs applying methods differing in costs, *Bioinformatics*, 23: 1519-1526.

A few results of the second part of the thesis have been used and cited in

> Bauer (2008): Adaptive designs: looking for a needle in the haystack - a new challenge in medical research, *Statistics in Medicine*, to appear.

A short summary of the first part can also be seen in Zehetmayer, Goll, Bauer and Posch (2007): Step by Step: mehr Effizienz mit neuen Studiendesigns, *Biospektrum*, 7: 754-755.

The two topics of the thesis consider two problems that may arise in genomic or proteomic studies. Hence after an introduction to the general methodology the two topics are treated in two separate sections of the thesis. Each section has its own specific introduction to the topic and concluding remarks. The relevant literature for the whole thesis is added at the end.

Availability:

R-programs (R (2005)) concerning the first part of the thesis to calculate asymptotically optimal designs are available on:

http://statistics.msi.meduniwien.ac.at/index.php?page=ao2stage

These R-programs can also be seen in the appendix.

2 Multiple hypothesis testing

When a single null hypotheses H is tested, a Type I error, that is rejecting the hypotheses, when it is in fact true (a false positive decision) may occur. A standard approach is to specify an acceptable level α for the probability of a Type I error. Let $H = 0$ if the null hypotheses is in fact true, and $H = 1$ if the alternative holds. The control of a specified Type I error probability α can be achieved by choosing a critical value c_α such that $P(T \geq c_\alpha \mid H = 0) \leq \alpha$, where T is the corresponding test statistic for hypothesis H. The hypothesis H is rejected if $T \geq c_\alpha$.

If the hypothesis is accepted, although in fact the alternative holds, a Type II error occurs (a false negative decision). The probability of a Type II error is: $\beta = P(T < c_\alpha \mid H = 1)$.

Multiple testing refers to the testing of more than one hypothesis at the same time. For example in genomic or proteomic experiments thousands of hypotheses are tested simultaneously. Since the probability of at least one Type I error increases with the number of hypotheses, in such studies large multiplicity problems occur. Therefore, problems that arise from the multiplicity aspect are:

- defining a Type I error rate and
- developing powerful multiple testing procedures that control this error rate.

Table 2.1: **Possible outcomes after a multiple testing procedure**

Number of	not rejected	rejected	Total
True null hypotheses	U	V	$m\pi_0$
False null hypotheses	T	S	$m(1-\pi_0)$
Total	m-R	R	m

2.1 Error rates

Consider the problem of testing simultaneously m null hypotheses H_i, $i = 1, ..., m$ and denote by R the number of rejected hypotheses among the m hypotheses. Assume that there are $m\pi_0$ true null hypotheses among all m hypotheses. The proportion of true null hypotheses π_0 is an unknown parameter. Table 2.1 describes the various outcomes when applying a multiple testing procedure to perform m hypothesis tests (compare Benjamini and Hochberg (1995)). The number of rejected hypotheses R is an observed random variable and S (number of true positive decisions), T (number of false negative decisions), U (number of true negative decisions) and V (number of false positive decisions) are unobservable random variables.

In the literature numerous different error rates and procedures for the control of these error rates are described. In the following two important error rates for the multiple testing situation are specified:

- The **Family Wise Error Rate (FWER)** is defined as the probability of at least one Type I error:

$$FWER = P(V \geq 1). \tag{2.1}$$

- The **False Discovery Rate (FDR)** is the expected proportion of Type I errors among the rejected hypotheses:

$$FDR = \mathbf{E}\left[\frac{V}{R} \mid R > 0\right] P(R > 0) = \mathbf{E}\left[\frac{V}{\max(R, 1)}\right]. \tag{2.2}$$

The effect $\max(R, 1)$ in the denominator is to set $V/R = 0$ if $R = 0$ (compare Benjamini and Hochberg (1995)).

Beside the FWER and the FDR, e.g. the Per-Comparison Error Rate ($PCER = \mathbf{E}(V)/m$), the Per-Family Error Rate ($PFER = \mathbf{E}(V)$) as well as the positive False Discovery Rate ($pFDR = \mathbf{E}[V/R \mid R > 0]$) proposed by Storey (2003) are described in the literature.

A multiple testing procedure is said to control a particular Type I error rate at level α if this error rate is less than or equal to α.
There is a distinction between strong and weak control of a Type I error rate. Strong control refers to the control of the Type I error rate under any combination of true and false null hypotheses. In contrast, weak control refers to the control of the Type I error rate only under the complete null hypotheses, that is when all null hypotheses are in fact true ($\pi_0 = 1$). Weak control is unsatisfactory, because in reality, some null hypotheses may be true and others false, but the subset of true null hypotheses is unknown. Strong control ensures that the Type I error rate is controlled under the unknown combination of true and false null hypotheses.

Under the complete null hypotheses the FDR is equal to the FWER (see Benjamini and Hochberg (1995)). Therefore control of the FDR implies control of the FWER in the weak sense. If $\pi_0 < 1$, the FDR is smaller than or equal to the FWER. Thus, procedures that control the FWER are more conservative, that is, lead to fewer rejections than those controlling the FDR. If a procedure only controls the FDR, more Type I errors but less Type II errors occur and thus, the power of the procedure may be increased. In the long run there is always a fraction of at most α true null hypotheses among the rejected hypotheses.

Within the class of multiple testing procedures that control a given Type I error rate at an acceptable level α, one seeks for test procedures that maximize the power $(1-\beta)$, that is, minimize the Type II error rate (β). As with Type I error rates, the concept of power can be generalized in various ways when moving from single to multiple hypotheses testing. Three common power definitions are (compare Dudoit, Shaffer and Boldrick (2003)):

- the probability of rejecting at least one false null hypothesis:
  $P(S \geq 1) = P(T \leq m(1-\pi_0) - 1)$

- the average probability of rejecting all false null hypotheses:
  $E[S]/(m(1-\pi_0))$

- the probability of rejecting all false null hypotheses:
  $P(S = m(1-\pi_0)) = P(T = 0)$

A number of articles can be found, where the problem of multiple testing is investigated for classical single-stage designs (e.g. Shaffer (1995)) or especially for the microarray setting (e.g. Dudoit et al. (2003) or Tusher, Tibshirani and Chu (2001)). Futschik and Posch (2005), for example, studied the problem on deciding on the number of hypotheses to be considered in a multiple hypothesis testing framework when the overall number of observations that can be collected is fixed. They showed that the expected number of detected effects can be increased by randomly selecting a smaller number of hypotheses such that more observations for each hypothesis are available.

2.2 Procedures controlling the family wise error rate

As mentioned before, the Family Wise Error Rate is defined as the probability of making at least one Type I error (see formula (2.1)). Well known general procedures controlling the FWER are the Bonferroni and the Bonferroni-Holm procedure which will further be explained. There are also other procedures, like the Sidak procedure, which is closely related to the Bonferroni procedure but is slightly less conservative or Hochbergs procedure, which can be viewed as step-up analog of Holm's step-down procedure, since the ordered p-values are compared to the same critical values in both procedures (see also Dudoit et al. (2003)).

2.2.1 The Bonferroni procedure

This procedure rejects any hypothesis H_i with a p-value less than or equal to α/m. The Bonferroni procedure is a single-step procedure, that means that equivalent multiplicity adjustments are performed for all hypotheses, regardless of the ordering of the test statistics or p-values. Each hypothesis is evaluated using a critical value that is independent of the result of tests of other hypotheses. The Bonferroni adjustment provides strong control of the FWER at level α since the actual probability of rejecting at least one null hypothesis is less than the nominal FWER level α (see e.g. Dudoit et al. (2003)). However, the power of this multiple testing procedure decreases strikingly with an increasing number of hypotheses m.

2.2.2 The Bonferroni-Holm procedure

The Bonferroni-Holm procedure is a step-down procedure, that means the hypotheses that correspond to the smallest p-values or largest absolute test statistics are considered successively, with further tests depending on the outcomes of earlier ones. As soon as one fails to reject a null hypothesis, no further hypotheses are rejected. For the strong control of the FWER at level α, this method proceeds as follows:

1. Let $p_1 \leq ... \leq p_m$ denote the observed ordered p-values and let $H_1, ..., H_m$ denote the corresponding null hypotheses.

2. Calculate $\hat{k} = \min\{1 \leq j \leq m : p_j > \alpha/(m - j + 1)\}$.

3. If $\hat{k}$ exists, then reject the null hypotheses H_j, for $j = 1, ..., \hat{k} - 1$.
   Otherwise, reject all hypotheses.

The Bonferroni-Holm procedure is less conservative than the standard Bonferroni procedure.

2.3 Procedures controlling the false discovery rate

In the genomic or proteomic setting, where thousands of tests are performed simultaneously and only a small number of genes or proteins are expected to be differentially expressed, FDR controlling procedures present a promising alternative to FWER approaches. In such situations, controlling the FWER can lead to unduly conservative procedures. One may tolerate some Type I errors, provided their number is small in comparison to the number of rejected hypotheses. The FDR, that is the expected proportion of Type I errors among the rejected hypotheses (see formula (2.2)), offers a less strict multiple testing criterion than the FWER. This different approach to multiple testing was proposed by Benjamini and Hochberg (1995).

Two approaches to provide FDR controlling procedures are the following: One is to fix the acceptable FDR level beforehand, and find a data-dependent thresholding rule so that the FDR of this rule is less than or equal to the pre-chosen level. This is the approach taken by Benjamini and Hochberg (1995). Another is to fix the thresholding rule and form an estimate of the FDR whose expectation is greater than or equal to the true FDR over that significance region. This is the approach taken by Storey (2002). These two procedures are discussed in the following.

2.3.1 The Benjamini-Hochberg procedure

Benjamini and Hochberg (1995) derived the following step-up procedure for strong control of the FDR for independent test statistics. In contrast to step-down procedures, step-up procedures begin with the largest p-value. Benjamini and Hochberg proved that the following procedure controls the FDR at a pre-chosen level α when the p-values following the null distribution are independent and uniformly distributed. This method proceeds as follows:

1. Let again $p_1 \leq ... \leq p_m$ denote the observed ordered p-values corresponding to the hypotheses $H_1, ..., H_m$.

2. For the control of the FDR at level α calculate

$\hat{k} = \max\{1 \leq k \leq m : p_k \leq \frac{k}{m}\alpha\}$.

3. If $\hat{k}$ exists, then reject the null hypotheses H_j for $j = 1, ..., \hat{k}$ corresponding to $p_1 \leq ... \leq p_{\hat{k}}$. Otherwise, reject nothing.

It can be shown that the Benjamini-Hochberg procedure controls the FDR in the strong sense (see Storey, Taylor and Siegmund (2004)). Benjamini and Yekutieli (2001) proved that this procedure also controls the FDR when the test statistics have positive dependency on each of the test statistics corresponding to the true null hypothesis. They also proposed, referring to Hommel (1988), a simple conservative modification of the procedure, replacing $\alpha k/m$ with $\alpha k/(m\sum_{j=1}^{m}\frac{1}{j})$ in the second step, which provides FDR control under arbitrary dependence structures (see also Dudoit et al. (2003)).

The Benjamini-Hochberg procedure was originally introduced by Simes (1986) to weakly control the FWER when all p-values are independent, although it happens to provide strong control of the FDR.

2.3.2 Storey's procedure

As mentioned before, instead of fixing α and estimating the rejection region, Storey (2002) fixed the rejection region and then estimated the FDR. Storey's method uses information about π_0, which yields a less stringent procedure and more power, while maintaining strong control. Typically the power of a multiple test procedure decreases with increasing m. But the larger m, the more information about π_0 is obtained.

Again m identical hypothesis tests $H_1, ..., H_m$ are performed with independent test statistics $T_1, ..., T_m$. Let $H_i = 0$ when the null hypothesis i is true and $H_i = 1$ otherwise. It is assumed that the test statistics under the true null $T_i|(H_i = 0)$ and under the alternative hypothesis $T_i|(H_i = 1)$ are identically distributed. It is further assumed that the same rejection region is used for each test. Finally it is assumed, that the H_i are independent Bernoulli random variables with $P(H_i = 0) = \pi_0$ and $P(H_i = 1) = 1 - \pi_0 = \pi_1$. Let Γ be the common

rejection region for each hypothesis test. Then the FDR can be written as:

$$\begin{aligned} FDR = P(H = 0 \mid T \in \Gamma) &= \frac{\pi_0 P(T \in \Gamma \mid H = 0)}{\pi_0 P(T \in \Gamma \mid H = 0) + \pi_1 P(T \in \Gamma \mid H = 1)} \\ &= \frac{\pi_0 P(T \in \Gamma \mid H = 0)}{P(T \in \Gamma)} \end{aligned} \tag{2.3}$$

In the following hypotheses are rejected on the basis of independent p-values. For rejections based on p-values, all rejection regions are of the form $[0, \gamma]$ for some $\gamma \geq 0$. In terms of p-values the above result can be written as:

$$FDR(\gamma) = \frac{\pi_0 P(p \leq \gamma \mid H = 0)}{P(p \leq \gamma)} = \frac{\pi_0 \gamma}{P(p \leq \gamma)} \tag{2.4}$$

where p is the random p-value resulting from any test.

Since π_0 is an unknown parameter, it has to be estimated. Storey (2002) proposed the following conservative estimate of π_0:

$$\hat{\pi}_0(\lambda) = \frac{\sharp\{p_i > \lambda\}}{(1-\lambda)m} = \frac{W(\lambda)}{(1-\lambda)m} \tag{2.5}$$

for some well-chosen λ, where $p_1, ..., p_m$ are the observed p-values, and $W(\lambda) = \sharp\{p_i > \lambda\}$ is the number of observed p-values exceeding λ. For a small proportion of null hypotheses this estimator can be larger than 1, thus in this cases it is set to 1. The argument for the choice of the estimator $\hat{\pi}_0(\lambda)$ could be explained as follows: As long as each test has reasonable power the large p-values are most likely to come from the true null hypothesis. Therefore for a well chosen λ, it is expected, that $\pi_0(1-\lambda)$ of the p-values lie in the interval $(\lambda, 1]$, because the p-values under the true null hypotheses are uniformly distributed. Therefore $W(\lambda)/m \approx \pi_0(1-\lambda)$, where $\mathbf{E}[\hat{\pi}_0(\lambda)] \geq \pi_0$ when the p-values corresponding to the true null hypotheses are uniformly distributed.

There is an inherent bias-variance trade off in the choice of λ. When λ gets smaller, the bias of $\hat{\pi}_0$ gets larger, but the variance gets smaller. Choosing a larger λ reduces the bias at the cost of higher variance (Storey et al. (2004)). Therefore, λ can be chosen to try to balance this trade-off. Storey (2002) optimized the value for λ to minimize the mean squared error

of the estimate with bootstrap methods. However, simulations showed that when choosing a non-optimal λ the difference in their true mean-squared errors is not very drastic. For his calculations he used $\lambda = 0.5$.

It is now assumed that λ is fixed. An estimate of $P(p \leq \gamma)$ is:

$$\widehat{P}(p \leq \gamma) = \frac{\sharp\{p_i \leq \gamma\}}{m} = \frac{R(\gamma)}{m}$$

where $R(\gamma) = \sharp\{p_i \leq \gamma\}$. The estimate for the FDR is then given by:

$$\widehat{FDR}_\lambda(\gamma) = \frac{\hat{\pi}_0(\lambda)\gamma}{\widehat{P}(p \leq \gamma)} = \frac{W(\lambda)\gamma}{(1-\lambda)\max\{R(\gamma), 1\}} \tag{2.6}$$

If $\widehat{FDR}_\lambda(\gamma) > 1$ Storey suggest setting $\widehat{FDR}_\lambda(\gamma) = 1$.

Storey, Taylor and Siegmund (2004) proved that if the p-values corresponding to the true null hypotheses are independent and uniformly distributed, then for fixed $\lambda \in [0, 1)$:

$$\mathbf{E}[\widehat{FDR}_\lambda(\gamma)] \geq FDR(\gamma)$$

for all γ and $\pi_0 < 1$.

For the finite sample case, that means in the case of a small m, Storey et al. (2004) introduced the following modification $\hat{\pi}_0^*$ of the estimator for π_0:

$$\hat{\pi}_0^*(\lambda) = \frac{W(\lambda) + 1}{(1-\lambda)m} \tag{2.7}$$

for $\lambda \in [0, 1)$. Therefore, the estimate of $FDR(\gamma)$ for the finite sample case is:

$$\widehat{FDR}_\lambda^*(\gamma) = \begin{cases} \frac{\hat{\pi}_0^*(\lambda)\gamma m}{\max\{R(\gamma),1\}} & \text{, if } \gamma \leq \lambda \\ 1 & \text{, if } \gamma > \lambda \end{cases} \tag{2.8}$$

Note that Storey (2002) fixed a rejection boundary γ and proposed an estimator for the FDR. To perform a test controlling a pre-chosen FDR α, the largest γ has to be determined, such that $\widehat{FDR}_\lambda(\gamma) \leq \alpha$. For $\lambda = 0$ Storey's procedure for a pre-chosen FDR is equivalent to the Benjamini-Hochberg procedure. For $\lambda > 0$, the rejection boundary γ is larger compared to the Bejamini-Hochberg method and thus it may be more powerful.

3 Two-stage designs applying methods differing in costs

3.1 Introduction

In gene expression and proteomic studies we generally deal with large numbers of hypotheses, where only for a small fraction of the hypotheses noticeable effects exist. Due to limited resources, the number of observations per hypothesis in a conventional single-stage design is low which limits the power. It has been shown that two-stage (or multi-stage) designs are a good option to improve the power. In these sequential designs, early stages are used to screen for the promising hypotheses, which are further investigated in later stages.

Two common approaches of such two-stage designs can be found in the literature. In the first type the sample sizes for the first and the second stage are preplanned. The second stage data is only collected for the hypotheses selected after the first stage and thus the total number of observations across stages and hypotheses (e.g.: total costs or overall gene evaluations) is random. Miller, Galecki and Shmookler-Reis (2001) suggested a two-stage design which uses first stage data only for the selection of hypotheses. After the second stage the test decision was only based on the second stage data ("pilot design"). To control the Family Wise Error Rate (FWER) a Bonferroni correction was performed for the second stage test. Satagopan and Elston (2003) improved this procedure using a group sequential approach (see also Jennison and Turnbull (2000)). Here the observations pooled over both stages were used for the final test ("integrated design"). Given a fix overall sample size they

focused on determining a two-stage design that minimizes the total cost of the study such that the desired overall significance level is controlled and the power is close to the desired power of the corresponding single-stage approach.

Since the control of the FWER may lead to conservative procedures, Benjamini and Yekutieli (2005) used the FDR concept for a two-stage procedure. In the first stage they test the m null hypotheses using the Benjamini-Hochberg procedure (see Section 2.3.1) for a prespecified level q_1. In the second stage for the selected hypotheses again a Benjamini-Hochberg procedure is used, now for a different level q_2 than in the first stage. They proved that if the test statistics are positively regression dependent on each hypothesis from the subset corresponding to true null hypotheses and for q_1 and q_2 fixed in advance, the FDR of their two-stage procedure is less or equal to $m_0 q_1 q_2 / m$, where m_0 is the number of true null hypotheses.

Van den Oord and Sullivan (2003) discussed the pilot and the integrated design controlling the FDR for more than two stages. Their general idea was that at earlier stages they allow for a high FDR. Markers that very likely do not have real effects are excluded from further analysis. At a later stage it will be needed to specify a low FDR to reduce the false discoveries. They minimized the overall observations for an intended power while controlling a pre-specified level for the FDR. The pilot and integrated two-stage design were further investigated in Bukszár and Van den Oord (2006).

In the second type of two-stage designs it is assumed, that there is a limitation on resources, that means the overall number of observations is fixed but not the overall sample size. A certain fraction of these resources is spent in the first stage for screening and the remaining resources are used in the second stage. Since the number of selected hypotheses is random, in this approach the second stage sample size is random. In the procedure of Satagopan, Verbel, Venkatraman, Offit and Begg (2002) the final test is performed using the combined subjects from both stages. A small prefixed number of hypotheses showing the highest test statistics after the second stage test is rejected. Their aim was to maximize the power with respect to the proportion of hypotheses selected for validation in the second stage and to

the proportion of resources allocated for the first stage. They showed that their procedure is more powerful than a single-stage design. However, this procedure neither controls the FWER nor the FDR. Instead of fixing the total number of observations Satagopan, Venkatraman and Begg (2004) fixed the total number of sample size so that no sample-size reallocation can be performed. They showed for the same procedure as in Satagopan et al. (2002) that two-stage designs can produce a reduction in gene evaluations for a minimal loss of power as compared to the single-stage design.

Zehetmayer, Bauer and Posch (2005) proposed (optimal) two-stage designs for experiments with a large number of hypotheses and constraints on the total sample size which control the FDR. All hypotheses whose conventional univariate first stage p-values are below a certain common threshold are selected for the second stage. The final test decision is based on the observations pooled over both stages. Zehetmayer et al. (2005) also investigated optimal pilot designs where the final test is only based on the second stage data (see also her doctoral thesis: Zehetmayer (2006)). Pilot designs controlling the FWER are discussed in Ohashi and Clark (2005). Further comparisons between the pilot and the integrated design can also be seen in Skol, Skott, Abecasis and Boehnke (2006). Recently Zehetmayer, Bauer and Posch (2008) have shown what can be achieved in terms of power by increasing the number of stages when total costs are fixed (see also their references to the literature on multi-stage designs).
In all these proposals constant costs and effect sizes over stages have been assumed.

In the following we investigate this type of two-stage designs where the costs per measurement differ between the first and second stage (see also Wang, Thomas, Pe'er and Stram (2006)). In modern genetic studies, there is an increasing focus on using a less accurate assay in early stages and more accurate ones in later stages for cost reasons. For example, a quasi-quantitative, global LC-MS profiling proteomics experiment may underestimate the true effect size due to saturation or sensitivity effects inherent in these multiplexed assays, whereas a targeted, calibrated assay (e.g. ELISA) can show an effect size generally larger than the profiling study. This work has been motivated by a study that aimed to define marker profiles

to predict response to chemotherapy using a proteomic approach. A two-stage design was planned to find such predictors. Due to the shortage of resources a low-cost standard method (2D Gelelectrophoresis) was applied at the first stage to screen for possible predictors. At the second stage a more expensive (and hopefully more effective) method was planned to be used for the promising predictors selected at the first stage (Western Blot). It was to expensive to apply the improved method for all patients and all proteins.

We consider two scenarios: In the first scenario different costs per measurement may arise if the same method is applied at both stages but specific experimental devices have to be produced at higher costs per measurement for the selected markers at the second stage, e.g. special chips have to be produced. In contrast to Wang et al. (2006) who constructed designs minimizing the overall costs for a given FWER and power, we assume that the total costs of the experiment are fixed, similar to Satagopan et al. (2002), Zehetmayer et al. (2005) or Ohashi and Clark (2005). We further consider in a second scenario that the experimenter from the beginning may have the choice between two methods that differ in costs and effect sizes. For limited total costs we derive both integrated and pilot designs with an asymptotically optimal power (for an increasing number of null hypotheses), either controlling the FWER or the FDR.

The test problem is defined in Section 3.2, the corresponding single-stage procedures in Section 3.3. In Sections 3.4 and 3.5 we define the asymptotically optimal pilot and integrated design controlling the FWER or the FDR. In Section 3.6 we give examples for the first scenario, where the costs per observation differ between stages, but the effect size remains the same. Scenario 1 is further investigated in Section 3.7 calculating cost ratios between stages for which it is worthwhile to use (optimal) two-stage designs (see Section 3.7.1). We further look how design misspecifications in the planning phase would change the power of two-stage designs as compared to the standard single-stage design (Section 3.7.2). For the second scenario we show that depending on the cost and the effect ratios between the methods it is preferable either to apply the low-cost or the high-cost method on both stages (Section 3.8).

Section 3.9 shows results under less stringent distributional assumptions like the situation of unknown variances (Section 3.9.1) or correlated hypotheses (Section 3.9.2). Results for the constraint of integer stage-wise sample sizes are given in Section 3.9.3. A discussion of all results is given in Section 3.10.

3.2 Test problem

Consider m_1 (null) hypotheses for the mean of independent normally distributed observations with known variance:

$H_{0i} : \mu_i = 0$ against $H_{1i} : \mu_i > 0$, $i = 1, ..., m_1$.

For deriving the test procedures, we assume independence of observations across hypotheses.

3.3 The single-stage design

We assume there is a limit for the overall total costs (resources) C of the study. Without loss of generality the costs per observation of the single-stage design are set to 1. In the standard single-stage design we equally allocate $n = C/m_1$ observations to each of the m_1 hypotheses. The test statistics used for decisions are the p-values $p_i = 1 - \Phi(z_i)$, $i = 1, ..., m_1$, where z_i is the standardized mean of the sample taken to test H_{0i} and Φ is the distribution function of the standard normal distribution. The p-values are compared to a common critical boundary γ: If $p_i < \gamma$ the null hypothesis H_{0i} is rejected, otherwise it is accepted. We further assume that for a fraction π_0 of the m_1 hypotheses considered the null hypothesis is true. To simplify later calculations we also assume that the same mean $\mu_i = \Delta\sigma$ holds true for all the alternatives, where σ^2 is the common known variance.

To control the FWER (the probability to reject at least one true null hypothesis irrespective of how many and which are in fact true, see Section 2.1) we apply the critical Bonferroni boundary $\gamma = \alpha/m_1$ (see Section 2.2). The power of such a single-stage design is defined by

$$\prod_s = 1 - \beta(\gamma) = 1 - \Phi_{\sqrt{\frac{C}{m_1}}\Delta,1}(c_{1-\gamma}),$$

where $\beta(\gamma)$ denotes the Type II error as a function of the rejection boundary γ, Φ_{μ,σ^2} is the distribution function of the normal distribution with mean μ and variance σ^2, and $c_{1-\gamma}$ is the $(1-\gamma)$-quantile of the standard normal distribution. Note that under the assumption of a common alternative the power is the expected fraction of null hypotheses correctly rejected.

To control the FDR (the expected proportion of erroneous rejections among all rejections, see Section 2.1) we apply the method of Storey (2002) estimating the FDR (see Section 2.3). The critical value γ is then determined as the maximum such that

$$\frac{\hat{\pi}_0 \gamma m_1}{\max(\sharp\{p_i < \gamma\}, 1)} \leq \alpha. \tag{3.1}$$

Here $\hat{\pi}_0$ is the estimated proportion of true null hypotheses given by

$$\hat{\pi}_0 = \sharp\{p_i > \lambda\}/\{(1-\lambda)m_1\}, \tag{3.2}$$

where $\lambda, 0 < \lambda < 1$, is a constant chosen a priori and $\sharp\{p_i > \xi\}$ denotes the number of p-values exceeding ξ. Hence the critical boundary is determined from the sample such that the estimated FDR never exceeds the targeted value α. Using the method of Storey the critical boundary is a random variable. Asymptotically (for $m_1 \to \infty$ and $C = d m_1$ for $d > 0$), γ can be determined from the equation

$$\alpha = \frac{\pi_0 \gamma}{\pi_0 \gamma + (1-\pi_0)(1-\beta(\gamma))}. \tag{3.3}$$

and plugged into the formula for $\prod_s = (1 - \beta(\gamma))$ to approximate the real power.

3.4 The pilot design

3.4.1 The test procedure

We consider the same test problem as described in Section 3.2. Again we assume there is a limit of overall total costs C for the study. Now a fraction r of the total costs C is used for the first stage for testing the m_1 hypotheses. Thus, for balanced sample size allocation the sample size of the first stage per hypothesis is

$$n_1 = rC/m_1.$$

The first stage p-values are given by $p_i^{(1)} = 1 - \Phi(z_i^{(1)})$ where $z_i^{(1)}$ is the first stage mean of the observations for hypothesis H_{0i}, $i = 1, ..., m_1$, standardized by using the common known first stage standard deviation σ_1. All null hypotheses are selected, whose p-values fall below a threshold γ_1 ($p_i^{(1)} < \gamma_1$). All others are accepted. Hence a random number of m_2 hypotheses is selected for the second stage.

Assume the sampling costs vary between the two stages due to producing specific experimental devices or applying a high-cost method in the second stage, so that the total costs are $m_1 n_1 + m_2 n_2 c_2 = C$ for some constant $c_2 \geq 1$. The remaining costs $(1 - r)C$ again are equally allocated over the selected null hypotheses so that the second stage sample size n_2 is given by

$$n_2 = \frac{C - m_1 n_1}{m_2 c_2} = \frac{(1 - r)C}{m_2 c_2}.$$

Let $z_i^{(2)}$ denote the mean of the second stage sample for hypothesis H_{0i}, now standardized by using the common known second stage standard deviation σ_2. Consequently $p_i^{(2)} = 1 - \Phi(z_i^{(2)})$ denotes the second stage p-value for the selected null hypothesis H_{0i}. In the pilot design the p-value used for decisions after the second stage is only calculated from the second stage sample. A selected hypothesis H_{0i} is rejected if the second stage p-value falls below the boundary γ_2 ($p_i^{(2)} < \gamma_2$). Otherwise it is accepted.

3.4.2 The pilot design controlling the FWER

To control the FWER we simply apply the Bonferroni method to determine the rejection boundary for the second stage p-value $p_i^{(2)}$, but in contrast to the single-stage design, the adjustment refers to the number of selected hypotheses m_2:

$$\gamma_2 = \alpha / m_2.$$

Since m_2 is independent of the second stage data, clearly this procedure controls the FWER at the level α.

We now will try to determine a γ_1 and r which maximizes the power of the two-stage design controlling the FWER. We assume that at stage one for all alternative hypotheses the same mean $\mu_{1i} = \Delta\sigma_1$ and at stage two the same mean $\mu_{2i} = k\Delta\sigma_2$, $k \geq 1$, holds true respectively. The advantage when using an improved method as compared to using the low-cost standard method is measured in terms of effect size, i.e. there may be a larger mean or a smaller variance when using the high-cost method. k is the ratio of the effect sizes between the two stages and we assume that the high-cost method at the second stage never provides a smaller effect size than the low-cost method at stage one. The first stage power for a true alternative is given by

$$1 - \beta_1(\gamma_1) = P_{\mu_{1i}=\Delta\sigma_1}(p_i^{(1)} < \gamma_1) = 1 - \Phi_{\sqrt{n_1}\Delta,1}(c_{1-\gamma_1}).$$

Note that under the assumption of a common alternative this is the expected proportion of correctly selected null hypotheses among all null hypotheses for which the alternative holds.

For the second stage we select a number of m_2 hypotheses which asymptotically is given by

$$m_2 = m_1(1 - \pi_0)(1 - \beta_1(\gamma_1)) + m_1\pi_0\gamma_1.$$

Because of the independence between the two stages, the overall power of the pilot design, that is the expected fraction of null hypotheses correctly rejected after the second stage among all null hypotheses for which the alternative holds, is asymptotically given by

$$\begin{aligned}\prod_p &= (1-\beta_1(\gamma_1))(1-\beta_2(\gamma_2)) \\ &= (1-\Phi_{\sqrt{n_1}\Delta,1}(c_{1-\gamma_1}))(1-\Phi_{\sqrt{n_2}\Delta k,1}(c_{1-\frac{\alpha}{m_2}})) \\ &= (1-\Phi_{\sqrt{\frac{rC}{m_1}}\Delta,1}(c_{1-\gamma_1}))(1-\Phi_{\sqrt{\frac{(1-r)C}{m_2 c_2}}\Delta k,1}(c_{1-\frac{\alpha}{m_2}})). \qquad (3.4)\end{aligned}$$

Given a FWER α, an initial number of hypotheses m_1, overall costs C, the cost ratio c_2 between stages, the proportion of true null hypotheses π_0, the effect size Δ and the effect size ratio k between stages we can optimize $\prod_p$ in the two design parameters r and γ_1. Considering r as an continuous variable the optimal sample sizes per stage (n_1 and n_2) in general will be non-integer. The case of integer stage-wise sample sizes is discussed later in Section 3.9.3. It is easy to see that the optimal γ_1 and r depend on C, m_1, Δ and k via $\sqrt{\frac{C}{m_1}}\Delta$ and $k/\sqrt{c_2}$ and the critical boundary α/m_2.

3.4.3 The pilot design controlling the FDR

To control the False Discovery Rate the second stage critical boundary γ_2 is determined as in formulas (3.1) and (3.2) replacing m_1 by m_2.

Asymptotically, for large m_1, the first stage selection boundary γ_1 and the second stage rejection boundary γ_2 in the pilot design have to adhere to the equation

$$\alpha = \frac{\pi_0\gamma_2\gamma_1}{\pi_0\gamma_2\gamma_1 + (1-\pi_0)(1-\beta_1(\gamma_1))(1-\beta_2(\gamma_2))} \qquad (3.5)$$

where $(1-\beta_1(\gamma_1))(1-\beta_2(\gamma_2))$ is the power $\prod_p$ of the pilot design defined in (3.4) using γ_2 instead of α/m_2:

$$\begin{aligned}\prod_p &= (1-\beta_1(\gamma_1))(1-\beta_2(\gamma_2)) \\ &= (1-\Phi_{\sqrt{\frac{rC}{m_1}}\Delta,1}(c_{1-\gamma_1}))(1-\Phi_{\sqrt{\frac{(1-r)C}{m_2 c_2}}\Delta k,1}(c_{1-\gamma_2})). \qquad (3.6)\end{aligned}$$

Again the power $\prod_p$ can be optimized as function of r and γ_1, where γ_2 follows from condition (3.5).

3.5 The integrated design

3.5.1 The test procedure

We address the same test problem as in Section 3.2. The screening step of the test procedure at the first stage is identical to the pilot design in the previous Section 3.4. The only difference to the pilot design is, that the final test decisions based on the selected null hypotheses are derived from integrated p-values $p_i = 1 - \Phi(z_i)$ which are based on the data from both stages. An obvious way to construct single combination test statistics z_i from both stages is to combine the stage-wise standardized means by suitable weights as applied for adaptive multi-stage clinical trials (see e.g. Lehmacher and Wassmer (1999)):

$$z_i = \sqrt{w_1} z_i^{(1)} + \sqrt{1 - w_1} z_i^{(2)}. \tag{3.7}$$

Now the test decision is again very simple: A selected null hypothesis H_{0i} is rejected in the final test if $p_i < \gamma$. Otherwise it is accepted. Optimizing the non-centrality parameter $(\sqrt{w_1}\sqrt{n_1} + \sqrt{1 - w_1}\sqrt{n_2}k)\Delta$ of the test statistics z_i leads to the optimal weight

$$w_1 = \frac{n_1}{n_1 + n_2 k^2}. \tag{3.8}$$

If the same method (with the same effect size, $k = 1$) is used at both stages then the weight $w_1 = n_1/(n_1 + n_2)$ corresponds to that used in a group sequential two-stage design. Note that using non-optimal weights may lead to a larger power of the pilot design as compared to the integrated design when the effect size in the second stage is much larger than in the first stage (as already pointed at by Skol et al. (2006)).

3.5.2 The integrated design controlling the FWER

Assuming n_2 as deterministic, the local level of the two-stage design γ when controlling the FWER is the solution of:

$$\gamma_s = P_{H_{0i}}(p_i^{(1)} < \gamma_1, p_i < \gamma) = \int_{c_{1-\gamma_1}}^{\infty} \left[1 - \Phi\left(\frac{c_{1-\gamma} - \sqrt{w_1} z}{\sqrt{1 - w_1}}\right)\right] \varphi(z) dz \tag{3.9}$$

where γ_s is set to α/m_1. φ denotes the density function of the standard normal distribution. Note again that in the test procedure described, n_2 is random because it depends on the number of selected hypotheses (which also is random). It follows from the assumption of independence of the observations across hypotheses, that the conditional distribution of n_2, given that the i-th hypothesis is selected, is independent of $p_i^{(1)}$. Thus the level (3.9) can be also used for random n_2.

By re-formulating the test decisions in terms of a sequential p-value p_{si} based on the monotonic ordering by Tsiatis, Rosner and Metha (1984) (H_{0i} is rejected if $p_{si} < \gamma_s$) one can show that the procedure mentioned above with the predefined sample size reallocation rule for the selected null hypotheses controls the FWER because under the null hypothesis they follow a uniform distribution (see Zehetmayer et al. (2005)):

The overall p-value for a group sequential two-stage test based on a monotonic ordering of the sample space as proposed in Tsiatis et al. (1984) is given by:

$$p_{si} = \begin{cases} p_i^{(1)} & \text{if } p_i^{(1)} > \gamma_1 \\ \int_{c_{1-\gamma_1}}^{\infty} \left[1 - \Phi(\frac{z_i - \sqrt{w_1} z}{\sqrt{1-w_1}})\right] \varphi(z) dz & \text{else} \end{cases} \tag{3.10}$$

(see Brannath, Bauer and Posch (2002)). The integral in (3.10) is the same as in (3.9) with $c_{1-\gamma_2}$ replaced by the observed z-statistics in the total sample. With (3.9), every critical region ($p_i^{(1)} \leq \gamma_1, p_i \leq \gamma$) corresponds to a critical region ($p_{si} \leq \gamma_s$) and vice versa. The p-value p_{si} is uniformly distributed under the null hypothesis H_{0i}. Thus, let γ_s ($0 \leq \gamma_s \leq 1$) be fixed and γ the solution of (3.9), then

$$P_{H_{0i}}(p_{si} \leq \gamma_s) = P_{H_{0i}}(p_i^{(1)} \leq \gamma_1, p_i \leq \gamma) = \gamma_s.$$

The overall power of an integrated two-stage design is given by

$$\begin{aligned} \prod_{int} &= P_{\mu_{1i}=\Delta\sigma_1, \mu_{2i}=k\Delta\sigma_2}(p_i^{(1)} < \gamma_1, p_i < \gamma) \\ &= \int_{c_{1-\gamma_1}}^{\infty} \left[1 - \Phi_{\sqrt{n_2}k\Delta,1}(\frac{c_{1-\gamma} - \sqrt{w_1} z}{\sqrt{1-w_1}})\right] \varphi_{\sqrt{n_1}\Delta,1}(z) dz \\ &= \int_{c_{1-\gamma_1}}^{\infty} \left[1 - \Phi_{\sqrt{\frac{(1-r)C}{m_2 c_2}}k\Delta,1}(\frac{c_{1-\gamma} - \sqrt{w_1} z}{\sqrt{1-w_1}})\right] \varphi_{\sqrt{\frac{rC}{m_1}}\Delta,1}(z) dz, \end{aligned} \tag{3.11}$$

where φ_{μ,σ^2} is the density function of the normal distribution with mean μ and variance σ^2. Given the other quantities C, c_2, m_1, π_0, Δ and k we can optimize $\prod_{int}$ in the two design parameters r and γ_1. Note that the optimal γ_1 and r, as in the pilot design, depend on C, m_1, Δ, k and c_2 via $\sqrt{\frac{C}{m_1}}\Delta$ and $k/\sqrt{c_2}$ and the critical boundary α/m_1.

As mentioned above, using non-optimal weights for the test statistics (3.7) leads to a larger power of the pilot design as compared to the integrated design when k is large. Figure 3.1 shows the asymptotic optimal power for varying k for the pilot design and the integrated design with and without optimal weights (see formula (3.8)). We consider the example of $m_1 = 1000$ hypothesis tests. A fraction of $\pi_0 = 0.99$ true null hypotheses and an effect size in the first stage of $\Delta = 0.5$ is assumed. Overall total costs are set to $C = 20000$ and the cost ratio is set to $c_2 = 15$. The targeted FWER is $\alpha = 0.05$. Figure 3.1 shows that whereas the integrated design with optimal weights with increasing k has a slightly larger power than the pilot design, the integrated design using non-optimal weights show a considerable loss of power as compared to the other two designs. For small k both integrated designs show

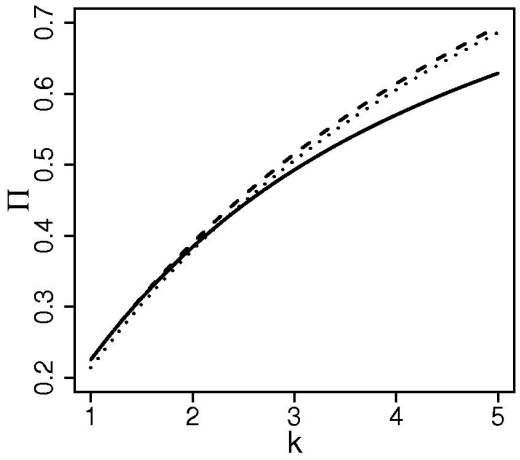

Figure 3.1: Asymptotically optimal power of the pilot design (dotted line) and the integrated design with (dashed line) and without (solid line) optimal weights for control of the FWER ($\alpha = 0.05$) for varying k:
$C = 20000$, $m_1 = 1000$, $c_2 = 15$, $\pi_0 = 0.99$, $\Delta = 0.5$.

nearly the same power and thus the power of both integrated designs is slightly larger then the power of the pilot design. For $k = 1$ the power values of the integrated designs with and without optimal weights are clearly the same. Further examples for the integrated design and the pilot design are considered later in Section 3.6.

3.5.3 The integrated design controlling the FDR

For the control of the False Discovery Rate, asymptotically the rejection boundary for the p-values in the final test is given by the solution of

$$\alpha = \frac{\pi_0 \gamma_s}{\pi_0 \gamma_s + (1 - \pi_0)(1 - \beta(\gamma_s))} \tag{3.12}$$

where γ_s is a function of γ which is given by (3.9). Such a two-stage procedure with a predefined sample size allocation rule controls the FDR since it can be shown that the resulting sequential p-values p_{si} are independent across hypotheses (see Zehetmayer et al. (2005)) such that the results of Storey (2002) on the consistency and conservativeness of the estimator of the FDR apply.

Again optimal values of r and γ_1 can be determined by maximizing the power (3.11) under the constraint (3.12). The rejection boundary γ for the p-values p_i of the selected null hypotheses, calculated from pooling stage-wise z-scores (3.7) with optimal weights (3.8), then can be found numerically from solving equation (3.9).

3.6 Examples: optimal designs for $k = 1$ and $c_2 \geq 1$

Asymptotically optimal two-stage designs applying the same method at both stages ($k = 1$) can be also derived in Zehetmayer et al. (2005) if the costs do not differ between stages ($c_2 = 1$) using appropriately defined total costs C. In the following we focus on designs using the same method at both stages, the second stage measurement however raising extra costs $c_2 > 1$. When $c_2 > 1$ we have to use the power formulas (3.4) (and (3.6) respectively when controlling the FDR) and (3.11) with $k = 1$ to derive asymptotically optimal designs.

Table 3.1 and 3.2 for $k = 1$ and some c_2 give the design parameters of optimal pilot and integrated designs and their power for controlling the FWER (Table 3.1) and the FDR (Table 3.2). Note that the optimal power values given for the integrated designs are only slightly larger than those of the pilot designs and both two-stage designs controlling the FDR have only a slightly larger power than when controlling the FWER. For comparison the power values of the (asymptotic) single-stage designs with equal total costs for the control of the FWER and FDR are listed in Table 3.3. As compared to the corresponding single-stage designs, the power of the two-stage designs is always considerably larger, even in the case of $c_2 = 15$.

As one can see from the tables, the asymptotic optimal screening boundary γ_1 decreases with increasing costs c_2 for the whole designs considered (pilot or integrated design controlling the FWER or FDR). For the same costs the screening boundary γ_1 slightly increases with increasing Δ. At the same time the proportion of costs used for the first stage r increases with Δ. Note that due to the complexity of the power function there is a different dependence of r on costs for low and large effect sizes, which is also depending on FDR or FWER control. At least the asymptotically optimal number of selected hypotheses m_2 increases with Δ and decreases with costs c_2 throughout the whole designs considered. The decrease of m_2 with increasing costs may be a consequence of the decreasing screening boundary γ_1.

Table 3.4 shows the asymptotic optimal design parameters for different values of the proportion of true null hypotheses π_0 for the case of $c_2 = 15$ and $\Delta = 0.75$. For increasing π_0, the

optimal fraction of total costs used in the first stage r is increasing, i.e. if the number of alternatives gets smaller, more costs and thus more sample size is needed in the first stage for screening to achieve the optimal power. There is a different dependence on the design used (pilot or integrated) and on FDR or FWER control for γ_1. With increasing π_0, the asymptotic optimal power is increasing and m_2 is decreasing .

Table 3.1: **Optimal two-stage designs controlling the FWER ($\alpha = 0.05$).** Asymptotically optimal parameters γ_1, r and the power ($\prod_p$ or $\prod_{int}$) as well as the second stage rejection boundary (γ_2 or γ) and m_2 for different c_2 and Δ. $k = 1$, $C = 20000$, $m_1 = 1000$, $\pi_0 = 0.99$.

Δ	c_2	Design	r	γ_1	$\prod_p$ or $\prod_{int}$	γ_2 or γ	m_2
0.5	1	pilot	0.635	0.07398	0.594	0.00063	79.55
		integrated	0.642	0.07665	0.603	0.00010	82.31
	5	pilot	0.683	0.01544	0.341	0.00262	19.06
		integrated	0.697	0.01599	0.351	0.00017	19.74
	15	pilot	0.685	0.00563	0.214	0.00621	8.04
		integrated	0.706	0.00611	0.226	0.00023	8.68
0.75	1	pilot	0.718	0.09209	0.926	0.00050	100.51
		integrated	0.725	0.10262	0.934	0.00006	111.03
	5	pilot	0.737	0.01870	0.762	0.00190	26.38
		integrated	0.759	0.02060	0.783	0.00009	28.50
	15	pilot	0.701	0.00686	0.589	0.00381	13.13
		integrated	0.745	0.00733	0.628	0.00010	14.03
1	1	pilot	0.774	0.09684	0.995	0.00047	105.83
		integrated	0.779	0.12025	0.997	0.00005	129.02
	5	pilot	0.781	0.01931	0.966	0.00173	28.82
		integrated	0.806	0.02368	0.974	0.00006	33.23
	15	pilot	0.722	0.00713	0.893	0.00309	16.17
		integrated	0.787	0.00825	0.925	0.00007	17.58

Table 3.2: **Optimal two-stage designs controlling the FDR ($\alpha = 0.05$).**
Asymptotically optimal parameters γ_1, r and the power ($\prod_p$ or $\prod_{int}$) as well as the second stage rejection boundary (γ_2 or γ) and m_2 for different c_2 and Δ. $k = 1$, $C = 20000$, $m_1 = 1000$, $\pi_0 = 0.99$.

Δ	c_2	Design	r	γ_1	$\prod_p$ or $\prod_{int}$	γ_2 or γ	m_2
0.5	1	pilot	0.632	0.09571	0.641	0.00356	101.57
		integrated	0.639	0.10070	0.651	0.00063	106.64
	5	pilot	0.701	0.01915	0.379	0.01053	23.17
		integrated	0.707	0.02024	0.387	0.00062	24.37
	15	pilot	0.716	0.00671	0.242	0.01922	9.44
		integrated	0.723	0.00707	0.249	0.00053	9.91
0.75	1	pilot	0.715	0.11907	0.943	0.00421	127.39
		integrated	0.722	0.13734	0.951	0.00063	145.57
	5	pilot	0.765	0.02455	0.810	0.01754	32.64
		integrated	0.778	0.02873	0.828	0.00069	37.00
	15	pilot	0.766	0.00871	0.673	0.04104	15.74
		integrated	0.788	0.01045	0.700	0.00064	17.82
1	1	pilot	0.772	0.12115	0.997	0.00437	129.91
		integrated	0.776	0.15811	0.998	0.00058	166.52
	5	pilot	0.807	0.02459	0.977	0.02111	34.15
		integrated	0.824	0.03296	0.983	0.00065	42.50
	15	pilot	0.799	0.00859	0.936	0.05793	17.97
		integrated	0.832	0.01209	0.954	0.00066	21.63

Table 3.3: **Single-stage designs controlling the FWER and FDR ($\alpha = 0.05$).**
Power $\prod_s$ for different Δ. $C = 20000$, $m_1 = 1000$ and $\pi_0 = 0.99$.

Δ	FWE-Control	FDR-Control
0.5	0.049	0.056
0.75	0.296	0.443
1	0.720	0.877

Table 3.4: **Optimal two-stage designs for different π_0.**
Asymptotically optimal parameters γ_1, r and the power ($\prod_p$ or $\prod_{int}$) as well as the second stage rejection boundary (γ_2 or γ) and m_2 for different π_0 for FWER or FDR Control: $\Delta = 0.75$, $c_2 = 15$ $k = 1$, $C = 20000$, $m_1 = 1000$.

	π_0	Design	r	γ_1	$\prod_p$ or $\prod_{int}$	γ_2 or γ	m_2
FWER	0.985	pilot	0.653	0.00680	0.543	0.00320	15.65
		integrated	0.718	0.00709	0.593	0.00010	16.75
	0.99	pilot	0.701	0.00686	0.589	0.00381	13.13
		integrated	0.745	0.00733	0.628	0.00010	14.03
	0.995	pilot	0.759	0.00676	0.641	0.00495	10.07
		integrated	0.784	0.00747	0.668	0.00010	10.95
FDR	0.985	pilot	0.753	0.00676	0.662	0.05895	19.46
		integrated	0.783	0.01104	0.694	0.00090	22.14
	0.99	pilot	0.766	0.00900	0.673	0.04104	15.74
		integrated	0.788	0.01045	0.700	0.00064	17.82
	0.995	pilot	0.788	0.00871	0.684	0.02258	11.55
		integrated	0.800	0.00934	0.704	0.00035	13.01

3.7 When to use two-stage designs

3.7.1 Break even point in the cost-ratio

It has been shown that for large m_1 and constraints on the total costs the power of an asymptotic optimal two-stage design (see Tables 3.1 and 3.2) may be considerably larger than the power of the corresponding single-stage design (Table 3.3). Again the scenario is considered where the same method is applied at the two stages ($k = 1$) and the second stage measurement raises extra costs ($c_2 > 1$).

We investigate when it is more efficient in terms of asymptotic power to use a two-stage design as compared to the single-stage design, i.e. we tackle the problem by asking whether a cost ratio c_2^* exists, where the power of the single-stage and the two-stage design are the same. If the asymptotic power would be monotonically decreasing in c_2 for $c_2 > c_2^*$ the single-stage design would provide a larger power than the two-stage design.

The first important answer is that for the integrated design such a finite c_2^* does not exist, because for given C, m_1, Δ, k and α and $c_2 \to \infty$ the power of the asymptotic optimal integrated design converges to the power of the single-stage design applying the low-cost measurement method ($\lim_{c_2 \to \infty} r = 1$). Hence for the integrated approach theoretically the two-stage approach always pays off. However in practice, if the optimal second stage sample size gets too small, the two-stage design cannot be used.

For the pilot design the power converges to 0 as $c_2 \to \infty$. Hence for the pilot design in general such a break even point c_2^* between the two-stage and single-stage design exists. Figure 3.2 shows the power values of the integrated and the pilot design as well as the power of the single-stage design controlling the FWER or the FDR for varying c_2.

Figure 3.3 shows c_2^* for varying π_0 and Δ for the case of controlling the FWER or the FDR at $\alpha = 0.05$. Again C is set to 20000 and m_1 is set to 1000. The curves are fairly sim-

ilar for control of the FWER and the FDR, the break even point varying more when the FDR is controlled. For large effect sizes, the power of the single-stage design and the pilot design are close to 1, and consequently c_2^* is small. For decreasing effect sizes the break even point c_2^* is increasing. When the number of true alternatives decreases (π_0 increases) c_2^* increases. In both situations a smaller number of null hypotheses is selected for the second stage (with larger sample sizes n_2) so that we can afford higher costs for the selected hypotheses. Note that the power when controlling the FDR is always slightly larger than when controlling the FWER. If there is a relatively large proportion of alternatives with substantial effects, the break even point is smaller for controlling the FDR than the FWER: the single-stage design controlling the FDR then is noticeably more powerful than the single-stage design controlling the FWER. For decreasing Δ this advantage in power of the single-stage FDR design over the single-stage FWER design decreases whereas the optimal two-stage design controlling the FDR still has favorable properties as compared to the two-stage FWER design. Hence larger second stage costs can be afforded to achieve the same power as the corresponding single-stage design. This may lead to a crossing of the two corresponding curves. For explanation see also Figures 3.4, which show the power values of the pilot design for different c_2 as well as the power of the corresponding single-stage design assuming $\pi_0 = 0.99$ (Figure (A)) and $\pi_0 = 0.995$ (Figure (B)).

For the example of $\Delta = 0.75$, the break even point is at $c_2^* = 81.38$ for FWER and at $c_2^* = 76.89$ for FDR control assuming $\pi_0 = 0.99$, that means if $c_2 > 81.38$ (76.89), the single-stage power is larger than the two-stage power (see Figure 3.3). Hence the two-stage design is preferable even if the cost ratio between stages is fairly high. Assuming $\pi_0 = 0.985$, the break even point is at $c_2^* = 57.63$ (FWER) and $c_2^* = 45.68$ (FDR) and assuming $\pi_0 = 0.995$ at $c_2^* = 141.36$ (FWER) and $c_2^* = 168.69$ (FDR). For $\pi_0 = 0.985$ and 0.99 the break even point when controlling the FWER is larger than when controlling the FDR. For $\pi_0 = 0.995$ the breakeven point controlling the FDR is larger.

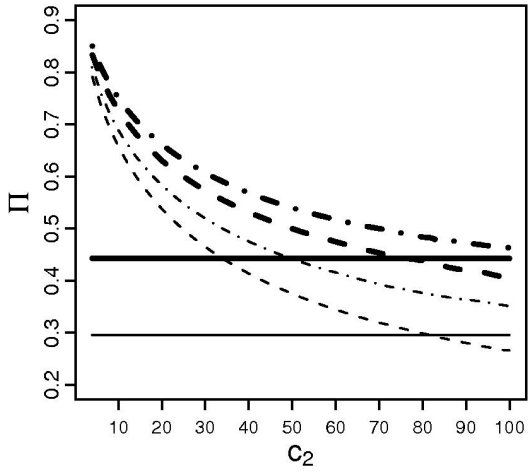

Figure 3.2: Power values for the integrated (dotdashed lines) and the pilot design (dashed lines) as well as the single-stage design (solid lines) controlling the FWER (thin lines) or FDR (bold lines) for varying cost ratio c_2. $C = 20000$, $m_1 = 1000$, $\Delta = 0.75$, $\pi_0 = 0.99$, FWER and FDR both $\alpha = 0.05$.

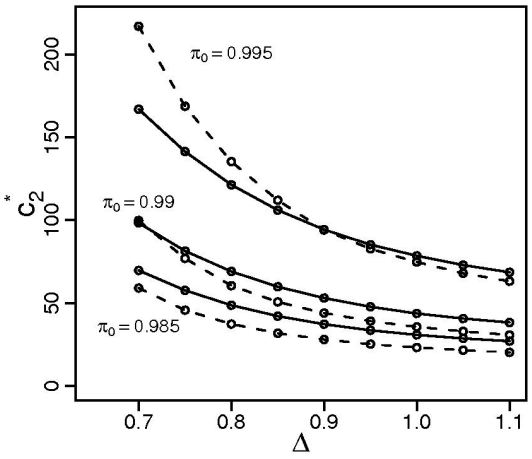

Figure 3.3: Break even point c_2^* for the cost ratio between the asymptotically optimal pilot design and the single-stage design depending on Δ and π_0, for controlling the FDR (dashed lines) or the FWER (solid lines).
$C = 20000$, $m_1 = 1000$, FWER and FDR both $\alpha = 0.05$.

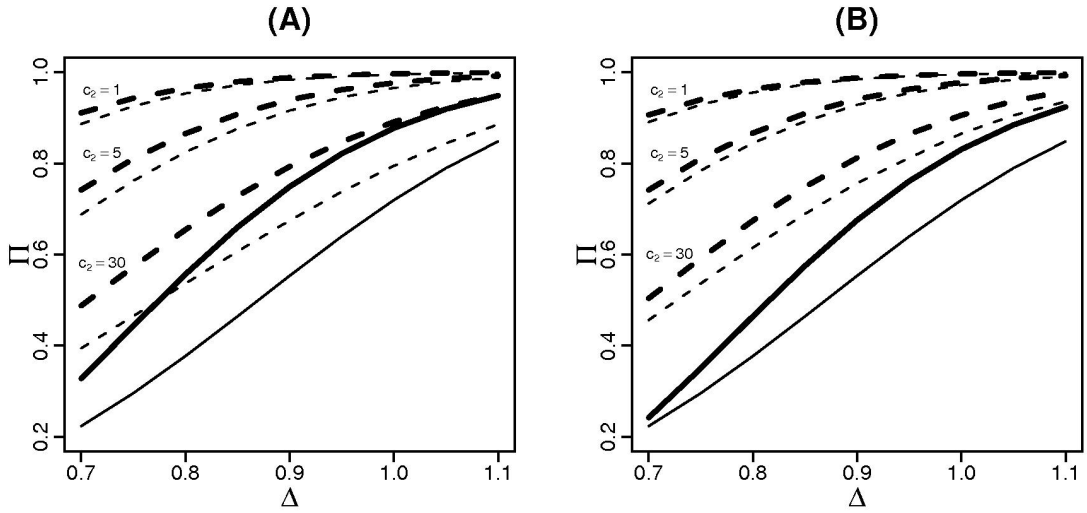

Figure 3.4: Power for the pilot design controlling the FWER (thin dashed lines) or FDR (bold dashed lines) as well as for the single-stage design (FWER: thin solid line, FDR: bold solid line) for different c_2 and Δ for $\pi_0 = 0.99$ (Figure (A)) and $\pi_0 = 0.995$ (Figure (B)). $C = 20000$, $m_1 = 1000$, FWER and FDR both $\alpha = 0.05$.

3.7.2 Impact of design misspecifications

Whereas costs are usually known a priori, the optimal designs depend on the unknown proportion of true null hypotheses π_0 and effect size Δ. Hence the impact of design misspecifications in the planning phase is an important issue.

We want to investigate the question whether there is an amount of misspecification where it would have been better to use a single-stage design. That means, whether we can misspecify the design parameters in that way, that the single-stage design would have a larger power than the two-stage approach. In the following again we consider the situation of $C = 20000$, $m_1 = 1000$ and $\alpha = 0.05$. It is assumed that the optimal r and γ_1 were planned for the situation where $\Delta = 0.75$, $\pi_0 = 0.99$ and $k = 1$. Figures 3.5 show the differences between the power of the two-stage designs and the corresponding single-stage designs as a function of the true π_0 and Δ for controlling the FDR and FWER for the example of $c_2 = 1$ and Figures 3.6 for $c_2 = 15$ (confer Wang et al. (2006)). Positive values indicate superiority of the two-stage design. The cross marks the point for which the two-stage design parameters were planned.

Figure 3.5 shows that assuming the same costs per observation in both stages (cost ratio $c_2 = 1$), for the shown parameter subspace for the true π_0 and Δ, the single-stage design neither outperforms the pilot design (first row of panels) nor the integrated design (second row) controlling the FWER (first column) or FDR (second column). One can also see from the figures, that the integrated design is more robust against misspecification of the design parameters than the pilot design. Note that for the pilot design such region, where it would have been better to perform a single-stage design exists, but only for severe design misspecifications (outside the parameter subspace shown). The power of the integrated design for $c_2 = 1$ is always larger or equal to the power of the single-stage design.

The example of fixing the cost ratio $c_2 = 15$ is plotted in Figure 3.6, again for the pilot (first row) and the integrated design (second row). Not surprising, the figures show that again the integrated design is more robust against misspecifications of π_0 and Δ than the pilot design: it uses the whole data set from both stages for test decisions. The most robust design is the integrated design controlling the FWER (Figure 3.6 (C)). Here in the parameter subspace shown the two-stage integrated design is always noticeably better than the single-stage design. Controlling the FDR, the advantage of the single-stage design to adapt for π_0 results in smaller differences between the integrated two-stage design and the single-stage design (Figure 3.6 (D)): in the left upper corner the single-stage design is outperforming the two-stage design. The bold line mark equality between the single-stage and the two-stage design.

The pilot design controlling the FWER is more sensible with regard to the design misspecifications than the pilot design controlling the FDR. The design applies "non-optimal" selection criteria and controlling the FWER no adaption to the correct parameters is possible in the second stage sample (Figure 3.6 (A)): in the left upper corner the power of the single-stage design may become substantially larger than the two-stage pilot design. Controlling the FDR adapting to the true parameters in the second stage sample helps a little (Figure 3.6 (B)): there is only a slightly larger power of the single-stage design as compared to the two-stage

pilot design in the left upper corner. Figures 3.7 show the power values for the pilot, the integrated and the single-stage design for control of the FWER (first column) or the FDR (second column), assuming that only one design parameter is misspecified, either Δ (first row) or π_0 (second row).

Generally, a design optimal for a fraction of true null hypotheses which is larger than the true π_0 can lead to a considerable loss of power as compared to the corresponding single-stage design. However, if the true π_0 gets larger than the proportion used for planning and the true effect size Δ is close to the one used for planning generally the difference between two-stage designs and the single-stage design increases. Optimism in the planning phase with regard to the number of true alternatives may help to avoid a loss of power due to design misspecification. If the true effect size Δ gets larger than the one from the planning phase for values of π_0 close to the true one the power of the two-stage and single-stage design both approach 1 so that the differences in the contour plots decrease.

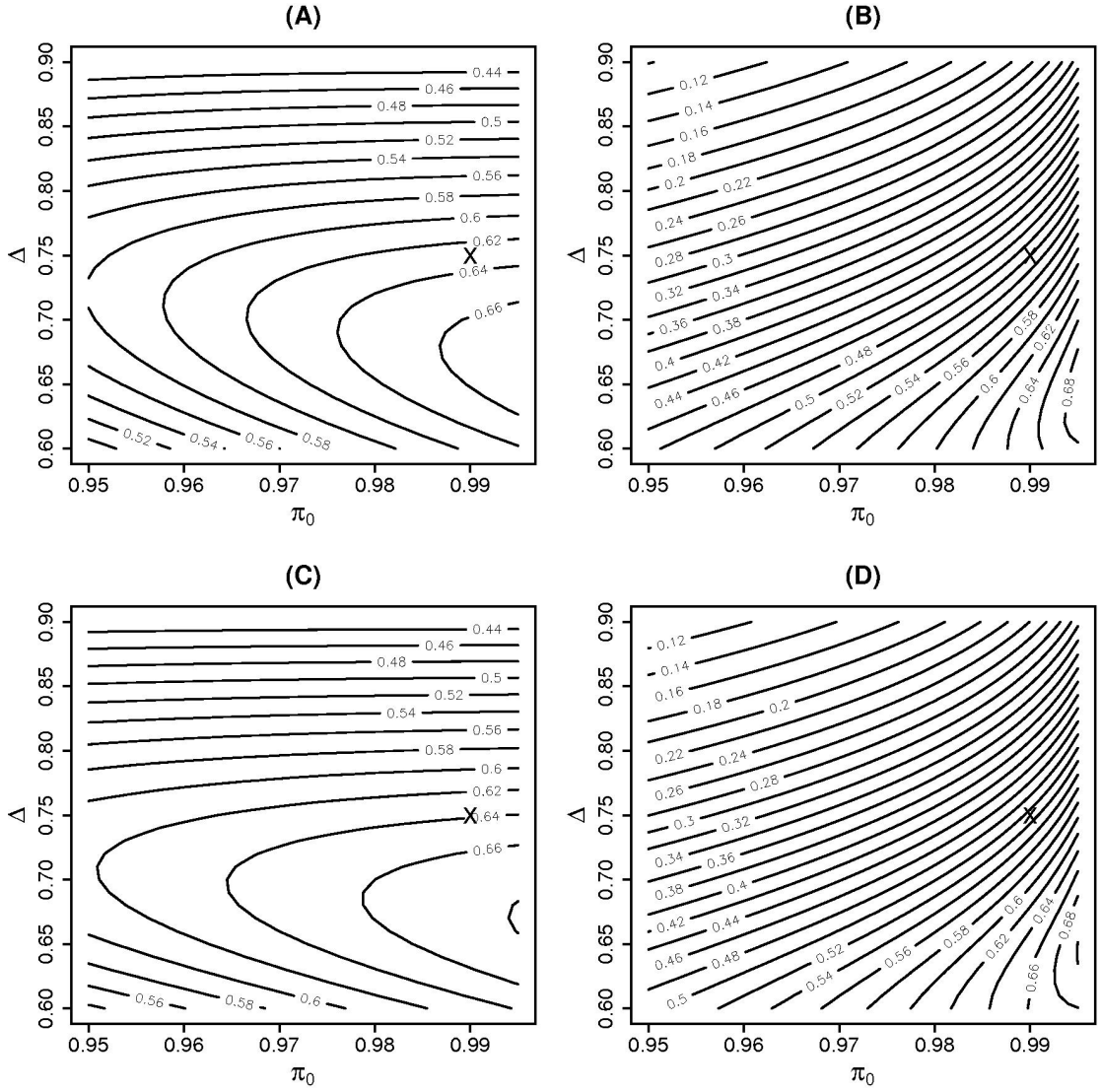

Figure 3.5: Contour plots for the difference in power between the single-stage and the pilot design (first row) and the single-stage and the integrated design (second row) as a function of the true π_0 and Δ for controlling the FWER (first column) or the FDR (second column). Positive values indicate superiority of the two-stage design. Asymptotically optimal two-stage designs were planned for $\pi_0 = 0.99$ and $\Delta = 0.75$ (marked as cross, confer Tables 3.1, 3.2 and 3.3).
$C = 20000$, $c_2 = 1$ and $m_1 = 1000$, FWER and FDR both $\alpha = 0.05$.

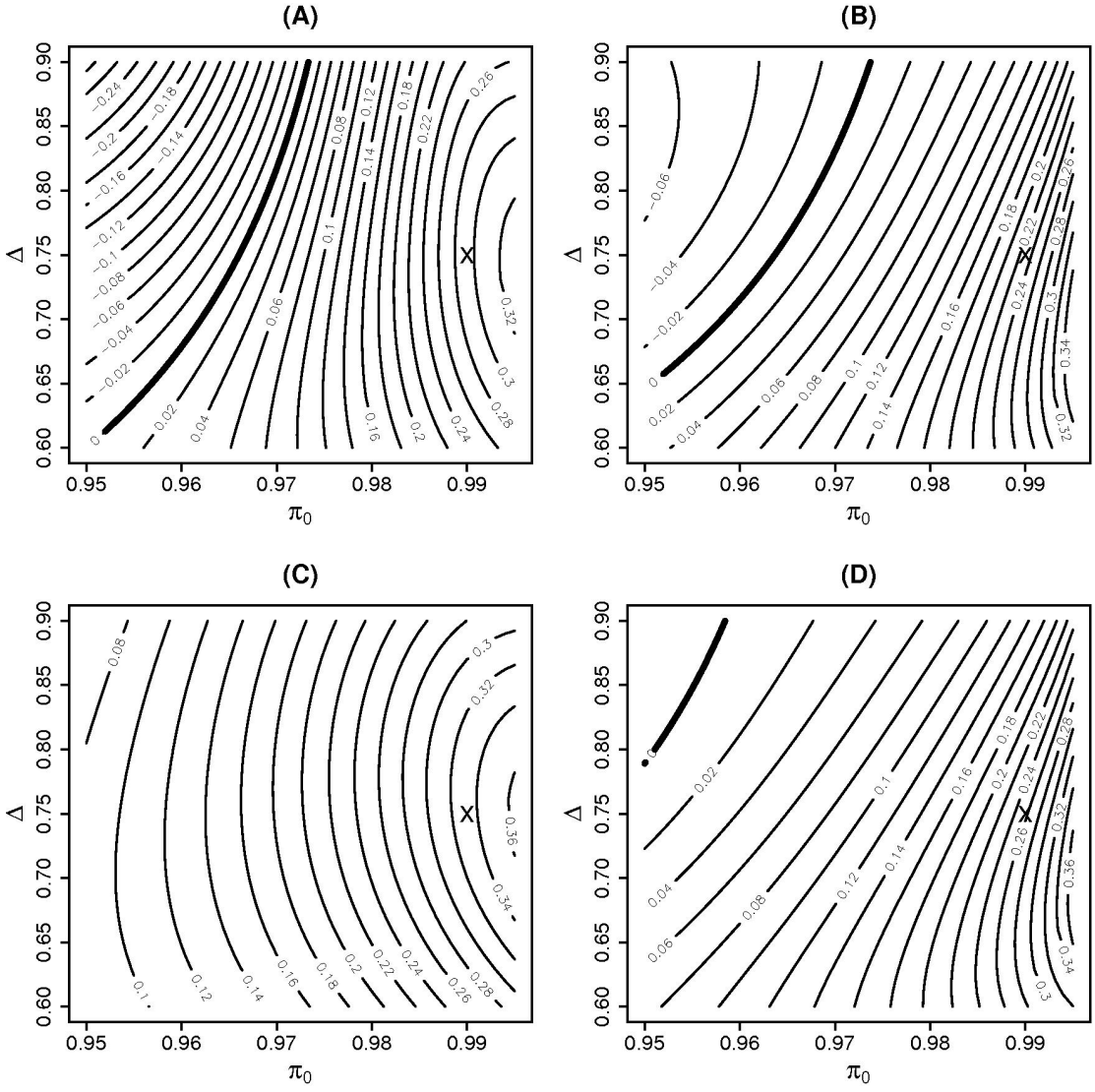

Figure 3.6: Contour plots for the difference in power between the single-stage and the pilot design (first row) and the single-stage and the integrated design (second row) as a function of the true π_0 and Δ for controlling the FWER (first column) or the FDR (second column). Positive values indicate superiority of the two-stage design. Bold lines mark equality between the single-stage and the two-stage design. Asymptotically optimal two-stage designs were planned for $\pi_0 = 0.99$ and $\Delta = 0.75$ (marked as cross, confer Tables 3.1, 3.2 and 3.3).
$C = 20000$, $c_2 = 15$ and $m_1 = 1000$, FWER and FDR both $\alpha = 0.05$.

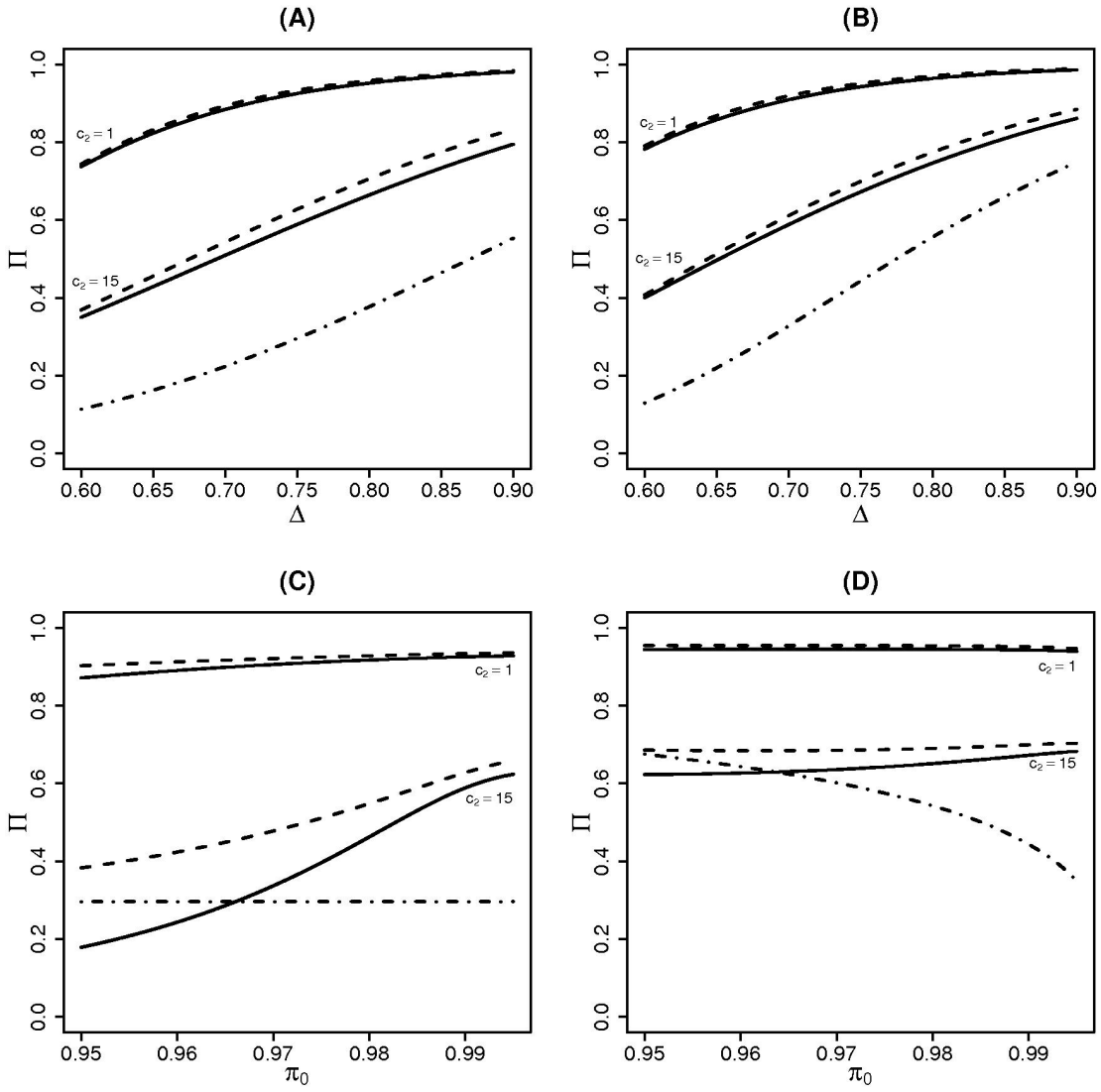

Figure 3.7: Power of the pilot design (solid lines), the integrated design (dashed lines) for $c_2 = 1$ and $c_2 = 15$ and the corresponding single-stage design (dotdashed lines) for varying true Δ (first row) and true π_0 (second row) controlling the FWER (first column) or the FDR (second column). Asymptotically optimal two-stage designs were planned for $\pi_0 = 0.99$ and $\Delta = 0.75$.
$C = 20000$ and $m_1 = 1000$, FWER and FDR both $\alpha = 0.05$.

3.8 Comparison of two-stage procedures

3.8.1 Two-stage procedures using the pilot design

We now assume that the experimenter has two different candidate methods for the measurements from the very beginning, a low-cost standard method and a high-cost improved method. So he could apply the same method at both stages (low-low or high-high) or he may switch to the more expensive method at the second stage (low-high).
In the following we investigate which of these three procedures is more powerful when controlling the FWER or the FDR. The power of the pilot design controlling the FWER for the low-high procedure is given by (3.4). Clearly the power of a procedure using the low-cost method in both stages, $\prod_{p_{ll}}$, say, is given by setting $k = 1$ and $c_2 = 1$ (see also Zehetmayer et al. (2005)):

$$\prod_{p_{ll}} = (1 - \Phi_{\sqrt{\frac{rC}{m_1}}\Delta,1}(c_{1-\gamma_1}))(1 - \Phi_{\sqrt{\frac{(1-r)C}{m_2}}\Delta,1}(c_{1-\frac{\alpha}{m_2}})). \tag{3.13}$$

The power for the procedure using the high-cost method at both stages,$\prod_{p_{hh}}$, say, arises from (3.4) by using $(1 - \Phi_{\sqrt{\frac{rC}{m_1 c_2}}k\Delta,1}(c_{1-\gamma_1}))$ for the first stage power leaving the second stage power unchanged:

$$\prod_{p_{hh}} = (1 - \Phi_{\sqrt{\frac{rC}{m_1 c_2}}k\Delta,1}(c_{1-\gamma_1}))(1 - \Phi_{\sqrt{\frac{(1-r)C}{m_2 c_2}}k\Delta,1}(c_{1-\frac{\alpha}{m_2}})). \tag{3.14}$$

It is easy to see that for $k = \sqrt{c_2}$ we get the identity $\prod_p \equiv \prod_{p_{ll}} \equiv \prod_{p_{hh}}$. Hence the maxima of all three functions in r and γ_1 are identical.

Since the power formulas for the three procedures are monotonic in c_2, the two-stage procedure applying the low-cost measurement method at both stages dominates the other two procedures (low-high and high-high) if the high-cost method is not sufficiently efficient, i.e., when $c_2 > k^2$. For $c_2 < k^2$ the high-high procedure dominates the other two. Hence the conclusion is that the procedure switching from the low-cost to the high-cost method is never the best procedure in terms of asymptotic power. However, it may be useful if the asymptotically optimal sample size at the first stage n_1 is too small for the high-high procedure, e.g.

in case of lack of finance so that the high-high procedure cannot be applied in the first stage.

Figures 3.8 show the maximum asymptotically optimal power over the three procedures for the pilot design (first row) for varying c_2, given the constraint $n_1 \geq 1$ (Figure 3.8 (A)) and $n_1 \geq 2$ (Figure 3.8 (B)) for the control of the FWER. Two different effect size ratios are assumed, $k = 3$ and 4. The example $C = 20000$, $m_1 = 1000$, $\pi_0 = 0.99$ and $\alpha = 0.05$ (FWER) is considered assuming an effect size for the low-cost measurement method of $\Delta = 0.5$. Thus the effect sizes of the high-cost method are assumed to be 1.5 and 2. The asymptotically optimal power is given for the three procedures (low-low: dotdashed line, low-high: dashed line, high-high: dotted line). The solid lines mark the respective maximal power over the three procedures if at least one (Figure 3.8 (A)) or if at least two (Figure 3.8 (B)) observations are left at the first stage for the optimal high-high procedure. Note that for the other two procedures the asymptotically optimal n_1 for the investigated values of c_2 is always larger than two. The power of the low-low procedure for the pilot design controlling the FWER is always $\prod_{p_{ll}} = 0.594$ since it does not depend on c_2 (compare also Table 3.1 in the Section 3.6).

Obviously the high-high procedure has the maximum power for relatively low costs c_2. For the cost ratio $k = 4$ the solid curve jumps when the costs of the high-cost method gets too large resulting in an asymptotic optimal $n_1 < 1$. This jump occurs at $c_2 = 13.11$ (Figure 3.8 (A)). Note that the crossing point with the low-low procedure is at $c_2 = 16$. Here the region where the low-high procedure is preferable to both the other is very small, for $k = 3$ no such region exists. If we apply the constraint $n_1 \geq 2$, the region where the low-high procedure is preferable gets larger, the jump between the high-high and the low-high procedure occurs at $c_2 = 7.17$ (Figure 3.8 (B)). For our example such a region does also exist for an effect ratio of $k = 3$, where the jump between the high-high and low-high procedure is at $c_2 = 6.66$. The common crossing point between the three procedures for $k = 3$ is at $c_2 = 9$.

For the control of the FDR the power of the pilot design for the low-high procedure (see formula (3.6)) has to be modified in the same way as for FWER control. Thus, the power for the low-low and the high-high procedure can be achieved by replacing α/m_2 in formulas

(3.13) and (3.14) by γ_2. Hence there is the same common crossing point for FDR control. Figures 3.9 (A) and (B) show the three procedures for the pilot design controlling the FDR. The figures look similar to them controlling the FWER. Note that the power for the low-low pilot procedure is always $\prod_{p_{ll}} = 0.641$ (compare Table 3.2 in Section 3.6). The jump between the high-high and low-high design is at $c_2 = 13.06$ for $k = 4$ under the constraint $n_1 \geq 1$ and at $c_2 = 7.15$ for $k = 4$ and $c_2 = 6.64$ for $k = 3$ under the constraint $n_1 \geq 2$. The area, where the low-high design is preferable, is only slightly larger for FDR control.

Note that in our example the high-cost method is much more effective than the low-cost method. If we have only double effects for the expensive method than the low-low procedure would be already preferable if the costs are only four times larger. Hence if the improved method is much more expensive it has to be much more effective to apply a high-high or a low-high procedure. Note also that the crossing point depends on the unknown effect size and no procedure dominates the other two over the whole parameter space. Hence in case of design misspecifications in the planning phase there will be other parameter constellations where the low-high type of strategy is in fact more powerful. However when no misspecifications occur, the low-high procedure may only be preferable if the high-cost method is too expensive so that the first stage sample size for the high-high procedure is insufficiently small.

In genomic or proteomic studies frequently there is no possibility for a low-low design, since methods for investigating selected hypotheses are more expensive than screening methods. For example the 2D-Gelelectrophoresis may be a good option for screening but it can not focus on special proteins so that no sample size reallocation can be performed. Thus it cannot be used in a second stage of a two-stage design. Therefore one may need a more improved method, like e.g. the Westernplot, which further investigates the 2D-Gelelektrophoresis. If it is not possible to use the low-low procedure, the low-high procedure may be a good option for a larger range of cost ratios c_2. As compared to the single-stage design, either using the high-cost method or using the low-cost method, the maximum over the three two-stage procedures for $c_2 > 1$ is larger, for small c_2 only slightly larger. For $c_2 = 1$ the power values of the high-cost single-stage design, the low-high and the high-high procedure in our example

are equal to one. Note that for the low-high pilot design for $c_2 = 1$ the optimal $\gamma_1 = 1$ and the optimal $r = 0$ and thus the optimal low-high pilot design is equal to the high-cost single-stage design. Figures 3.10 show again the optimal power of the three two-stage procedures for the pilot designs and additionally the power of the single-stage design using the high-cost method or using the low-cost method controlling the FWER (Figure (A)) and the FDR (Figure (B)) for varying $c_2 \geq 2$. Whereas the power of the low-cost single-stage design is always small, the power of the high-cost single-stage design for small c_2 is similar to those from the low-high procedure, and thus, larger than the power of the low-low procedure. For increasing c_2 the high-cost single-stage design considerably looses power and for large c_2 it falls below the power of the low-cost single-stage design.

When an experimenter from the beginning may have the choice between a low-cost and a high-cost method, there is also a fourth possibility of a two-stage procedure not mentioned above, to screen with the high-cost method and investigate the selected hypotheses with the low-cost method. In the genomic or proteomic studies, this combination in most cases does not make sense. As mentioned above, the methods that are able to look at selected hypotheses are often the expensive, improved methods and the low-cost methods are preferable for screening. Thus the high-low procedure often is not possible to apply. Nevertheless, we had a look at this combination for the two-stage procedures controlling the FWER. Figure 3.11 (A) again shows the asymptotically optimal power of the low-low, the low-high and the high-high procedure for varying c_2 and additionally the power of the high-low procedure is given. The solid line shows the maximal power over the four designs under the constraint of $n_2 \geq 2$ for the high-high and the high-low procedure. The area, were the high-low procedure is preferable is strikingly small. Applying the high-low design the optimal n_1 for the screening step is only slightly larger than applying the high-high design. Since we use the high-cost method in both designs for screening similar first stage sample sizes are needed. However, a much smaller m_2 is selected for the second stage leading to a much larger n_2 when using the low-cost method in the second stage. Assuming an effect size ratio of $k = 4$, this area is between $c_2 = 7.17$ and $c_2 = 7.36$ and assuming an effect size ratio of $k = 3$ the high-low procedure is preferable if $6.66 \leq c_2 \leq 6.77$.

3.8.2 Two-stage procedures using the integrated design

Comparing the three procedures for the integrated design we have to modify the formula for the power $\prod_{int}$ given for the low-high procedure in (3.11). For the low-low procedure to calculate the power we have to insert $k = 1$ and $c_2 = 1$ (see Zehetmayer et al. (2005)):

$$\prod_{int_{ll}} = \int_{c_{1-\gamma_1}}^{\infty} \left[1 - \Phi_{\sqrt{\frac{(1-r)C}{m_2}}\Delta,1}\left(\frac{c_{1-\gamma} - \sqrt{w_1}z}{\sqrt{1-w_1}}\right)\right] \varphi_{\sqrt{\frac{rC}{m_1}}\Delta,1}(z)dz. \tag{3.15}$$

For the high-high procedure we have to replace $\sqrt{n_1}\Delta$ by $\sqrt{\frac{n_1}{c2}}k\Delta$:

$$\prod_{int_{hh}} = \int_{c_{1-\gamma_1}}^{\infty} \left[1 - \Phi_{\sqrt{\frac{(1-r)C}{c_2 m_2}}k\Delta,1}\left(\frac{c_{1-\gamma} - \sqrt{w_1}z}{\sqrt{1-w_1}}\right)\right] \varphi_{\sqrt{\frac{rC}{c_2 m_1}}k\Delta,1}(z)dz. \tag{3.16}$$

It can be seen easily that again for $k = \sqrt{c_2}$ the three power functions are identical so that there is the same crossing point for the integrated design. Essentially the results are very similar to those for the pilot design (see Figure 3.8 (C) and (D) for the control of the FWER and Figure 3.9 (C) and (D) for the FDR control). Controlling the FDR the modification of the power formula is similar.

For $k = 4$ the jump between the high-high design and the low-high design under the constraint $n_1 \geq 1$ occurs at $c_2 = 13.18$ when controlling the FDR and $c_2 = 13.24$ for FWER control. No such jump exists for $k = 3$. Under the constraint $n_1 \geq 2$ for $k = 4$, this jump is at $c_2 = 7.20$ for FDR and at $c_2 = 7.24$ for FWER control. For $k = 3$, the jump occurs at $c_2 = 6.70$ for FDR and $c_2 = 6.73$ for FWER control. Note that the power of the low-low procedure is always 0.651 for FDR and 0.603 for FWER control (again compare Tables 3.1 and 3.2). The area where the low-high procedure is preferable, is slightly smaller as compared to the pilot design. Again, as compared to the two possible single-stage designs, as in the pilot design, the maximum over the three two-stage procedures is always larger.

Note that the common crossing point only exists if in the integrated low-high procedure the optimal weights (3.8) are used for combining the stage-wise test statistics. The low-high procedure loses power when applying non-optimal weights.

We also investigated the high-low procedure for the integrated design controlling the FWER under the constraint $n_1 \geq 2$ (see Figure 3.11 (B)). The results are similar to the pilot design. The area where this procedure is preferable is again very small, but slightly larger than for the pilot design. For $k = 4$ this procedure is preferable if $7.24 \leq c_2 \leq 7.47$ and for $k = 3$ if $6.73 \leq c_2 \leq 6.84$.

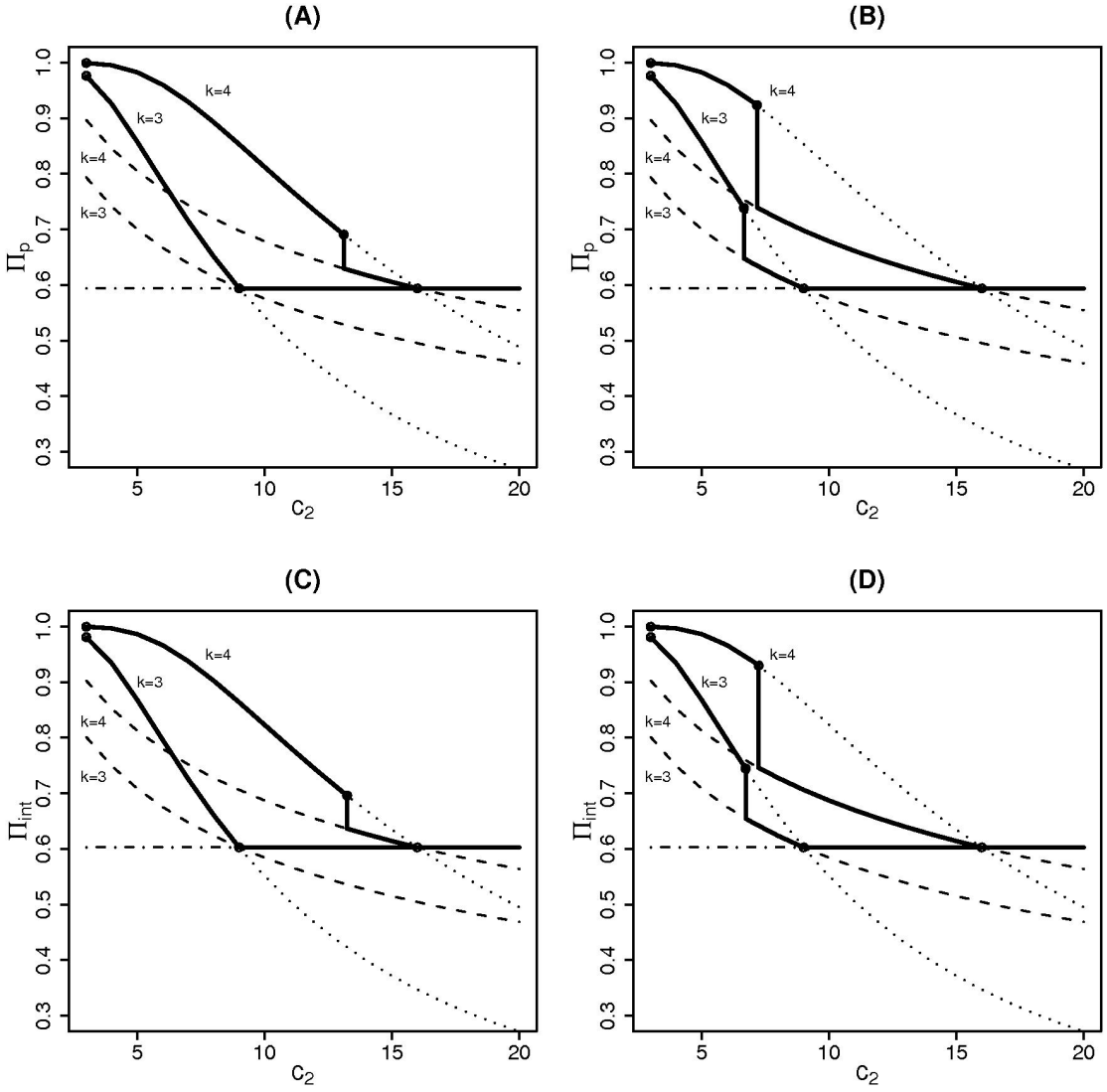

Figure 3.8: Asymptotically optimal power of the low-low (dotdashed horizontal line), the low-high (dashed lines) and the high-high (dotted lines) procedure of the pilot design (first row) and the integrated design (second row) controlling the FWER ($\alpha = 0.05$) for varying c_2 and effect size ratios $k = 3$ and $k = 4$. The solid lines mark the respective maximum over the three procedures under the constraint $n_1 \geq 1$ (first column) and $n_1 \geq 2$ (second column) for the high-high procedure. $C = 20000$, $m_1 = 1000$, $\pi_0 = 0.99$, $\Delta = 0.5$.

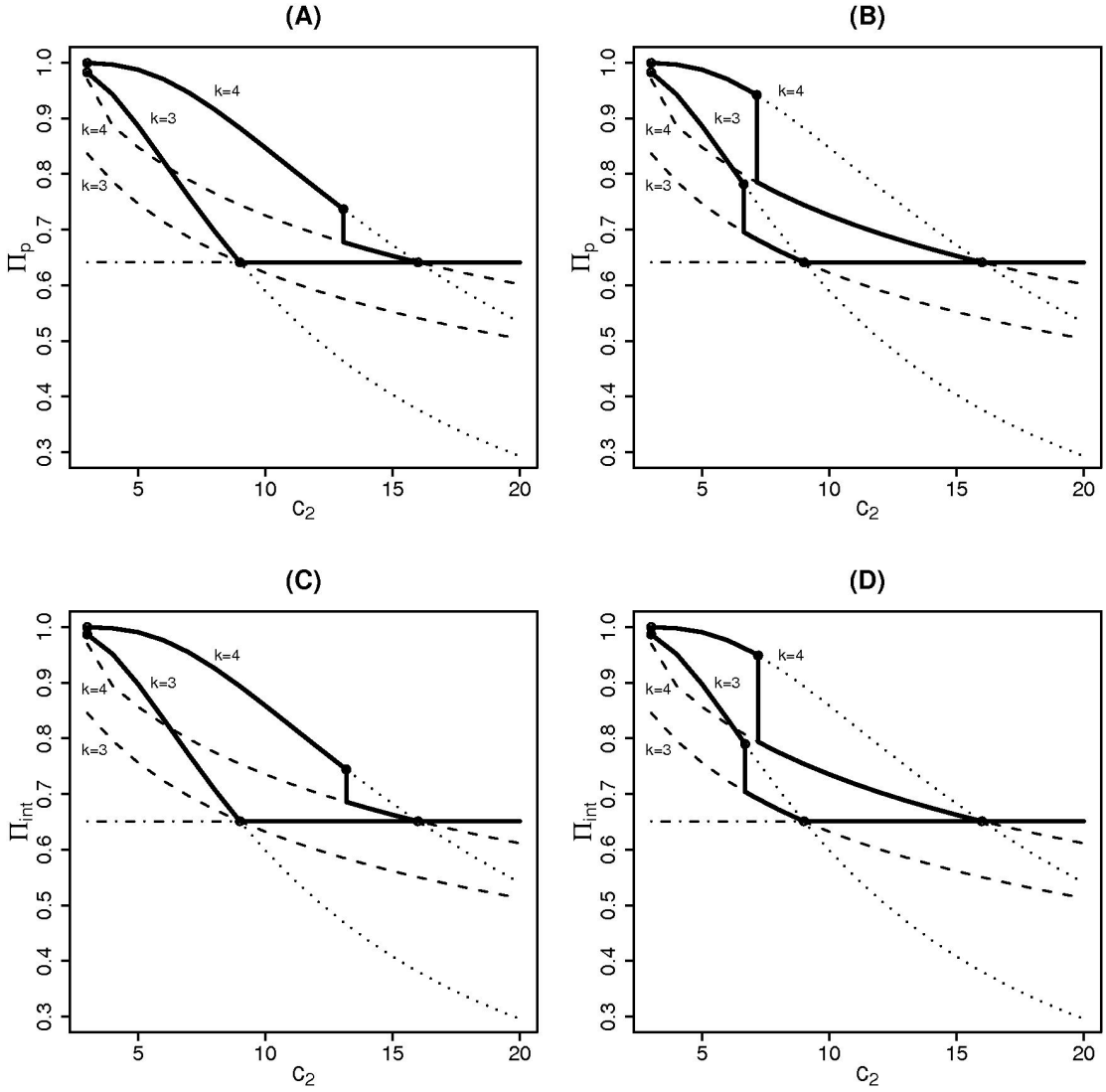

Figure 3.9: Asymptotically optimal power of the low-low (dotdashed horizontal line), the low-high (dashed lines) and the high-high (dotted lines) procedure of the pilot design (first row) and the integrated design (second row) controlling the FDR ($\alpha = 0.05$) for varying c_2 and effect size ratios $k = 3$ and $k = 4$. The solid lines mark the respective maximum over the three procedures under the constraint $n_1 \geq 1$ (first column) and $n_1 \geq 2$ (second column) for the high-high procedure.
$C = 20000$, $m_1 = 1000$, $\pi_0 = 0.99$, $\Delta = 0.5$.

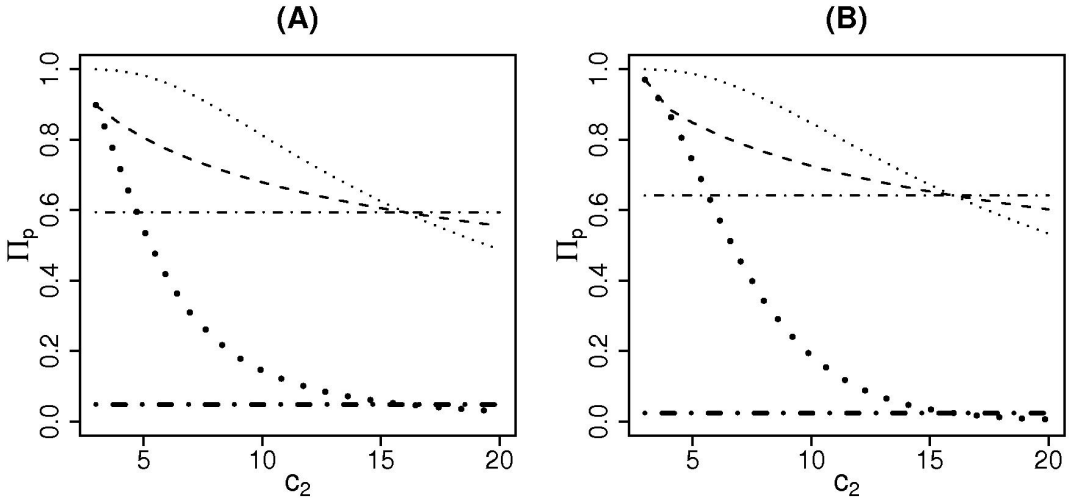

Figure 3.10: Asymptotically optimal power of the low-low (dotdashed horizontal line), the low-high (dashed lines), the high-high (dotted lines) as well as the single-stage design using the high-cost (bold dotted line) or the low-cost method (bold dotdashed line) for the pilot design controlling the FWER (Figure (A)) and the FDR (Figure (B))for varying c_2. $k = 4$, $C = 20000$, $m_1 = 1000$, $\pi_0 = 0.99$, $\Delta = 0.5$, FWER and FDR both $\alpha = 0.05$.

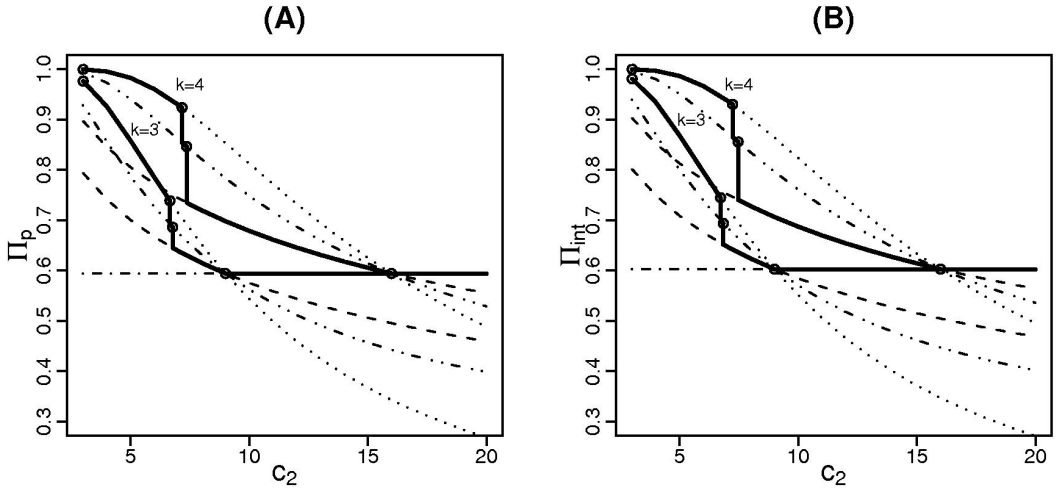

Figure 3.11: Asymptotically optimal power of the low-low (dotdashed horizontal line), the low-high (dashed lines), the high-high (dotted lines) and the high-low (dotdotdashed lines) procedure of the pilot design (Figure (A)) and the integrated design (Figure (B)) controlling the FWER ($\alpha = 0.05$) for varying c_2 and effect size ratios $k = 3$ and $k = 4$. The solid lines mark the respective maximum over the four procedures under the constraint $n_1 \geq 2$ for the high-high and the high-low procedure. $C = 20000$, $m_1 = 1000$, $\pi_0 = 0.99$, $\Delta = 0.5$.

3.9 Extensions

3.9.1 The situation of unknown variances

To investigate the situation of unknown variances we performed optimizations for the pilot design. Because of the independence between the first and the second stage, the pilot design can be easily applied to the unknown variance case by using the central and non-central t-distributions instead of the corresponding normal distributions. Thus, the power of the pilot design controlling the FWER for the unknown variance case is given by:

$$\prod_p^t = (1 - F_{\sqrt{\frac{rC}{m_1}}\Delta, n_1-1}(c^t_{1-\gamma_1, n_1-1}))(1 - F_{\sqrt{\frac{(1-r)C}{m_2 c_2}}\Delta k, n_2-1}(c^t_{1-\frac{\alpha}{m_2}, n_2-1})), \tag{3.17}$$

where $F_{\mu,n}$ is the distribution function of the non-central t-distribution with non-centrality parameter μ and n degrees of freedom and $c^t_{1-\gamma_1,n}$ is the $(1-\gamma_1)$-quantile of the central t-distribution with n degrees of freedom. The Power $\prod_p^t$ can now, as in the known variance case, be optimized as function of r and γ_1 for a given FWER α, an initial number of hypotheses m_1, overall costs C, a cost ratio c_2, a proportion of true null hypotheses π_0, an effect size Δ and an effect size ratio k.

For the control of the FDR the power of the pilot design for the unknown variance case can also be easily applied by replacing α/m_2 in formula (3.17) by γ_2:

$$\prod_p^t = (1 - F_{\sqrt{\frac{rC}{m_1}}\Delta, n_1-1}(c^t_{1-\gamma_1, n_1-1}))(1 - F_{\sqrt{\frac{(1-r)C}{m_2 c_2}}\Delta k, n_2-1}(c^t_{1-\gamma_2, n_2-1})). \tag{3.18}$$

Asymptotically (for $m_1 \to \infty$ and $C = dm_1$ for $d > 0$), the second stage rejection boundary γ_2 have to adhere to the equation

$$\alpha = \frac{\pi_0\gamma_2\gamma_1}{\pi_0\gamma_2\gamma_1 + (1-\pi_0)\prod_p^t}. \tag{3.19}$$

Again the power $\prod_p^t$ calculated with formula (3.18) can be optimized as function of r and γ_1 where γ_2 follows from condition (3.19).

Table 3.5: **Optimal pilot designs controlling the FWER or FDR ($\alpha = 0.05$) for the unknown variance case.**
Asymptotically optimal parameters γ_1, r as well as the number of selected hypotheses m_2 and the power using the optimal parameters from the unknown ($\prod_p^t$) and the known variance case ($\prod_p^{t*}$) for different c_2 and Δ. The power values $\prod_s$ of the corresponding single stage designs are given.
$k = 1$, $C = 20000$, $m_1 = 1000$, $\pi_0 = 0.99$.

	Δ	c_2	r	γ_1	m_2	$\prod_p^t$	$\prod_p^{t*}$	$\prod_s$
FWER	0.5	1	0.622	0.075	80.53	0.555	0.555	0.020
		5	0.681	0.016	19.08	0.283	0.282	
		15	0.692	0.006	7.97	0.162	0.160	
	0.75	1	0.696	0.094	102.34	0.905	0.903	0.122
		5	0.722	0.020	26.79	0.681	0.680	
		15	0.702	0.007	12.48	0.473	0.471	
	1	1	0.747	0.100	109.15	0.992	0.991	0.391
		5	0.755	0.021	30.48	0.932	0.931	
		15	0.709	0.008	16.10	0.791	0.791	
FDR	0.5	1	0.618	0.097	102.77	0.608	0.608	0.039
		5	0.695	0.020	22.92	0.318	0.318	
		15	0.718	0.007	8.82	0.180	0.180	
	0.75	1	0.694	0.123	130.70	0.930	0.929	0.159
		5	0.748	0.026	33.88	0.747	0.746	
		15	0.757	0.009	15.52	0.565	0.564	
	1	1	0.747	0.127	135.64	0.995	0.995	0.660
		5	0.784	0.028	37.28	0.957	0.957	
		15	0.783	0.010	19.08	0.877	0.875	

Table 3.5 shows the optimal parameters r and γ_1 for the pilot design for the unknown variance case controlling the FWER or FDR for different cost ratios c_2 and effect sizes Δ. The effect size ratio between stages k was set to 1. Again the example of $C = 20000$, $m_1 = 1000$ and $\pi_0 = 0.99$ is considered. The optimal power values using the optimal parameters for the unknown variance case ($\prod_p^t$) are shown. For comparison to the known variance case see Tables 3.1 and 3.2. The optimal power values in the unknown variance case clearly decrease as compared to the power values for the known variance case. The number of the selected hypotheses m_2, like in the known variance case, decreases with increasing c_2 and increases with increasing Δ. As compared to the known variance case always a smaller m_2 is selected for the second stage through the whole examples considered.

Table 3.5 also shows the power $\prod_p^{t*}$ using the optimal parameters from the known variance case. It can be seen that using the parameters of the known variance case in the situation of unknown variances leads to virtually the same performance as using the optimal parameters from the unknown variance case.

The impact of misspecification is investigated in Figures 3.12. Contour plots for the difference in power between the single-stage and the pilot design as a function of the true π_0 and Δ for controlling the FWER (first row) or the FDR (second row) for $c_2 = 1$ (first column) and $c_2 = 15$ (second column) are shown. Positive values indicate superiority of the two-stage design. Again the example of $C = 20000$ and $m_1 = 1000$ is investigated. Asymptotically optimal two-stage designs were planned for $\pi_0 = 0.99$ and $\Delta = 0.75$ (marked as cross, confer Table 3.5). The results are similar to the known variance case, for comparison see Figures 3.5 and 3.6. Assuming a cost ratio of $c_2 = 1$, in the parameter space shown, the single-stage design is not outperforming the pilot design (Figures (A) and (B)). For $c_2 = 15$ the single-stage design, as in the known variance case, in the left upper corner is outperforming the pilot design (Figures (C) and (D)).

Finally we will have a look at Scenario 2 for the unknown variance case. The decision which of the three two-stage procedures (low-low, high-high, low-high) is preferable, is more difficult because no common crossing point in costs as a function of c_2 between the three procedures exists. However, the region where the low-high procedure is preferable still remains small. Figures 3.13 shows the asymptotically optimal power for the unknown variance case of the low-low (dotdashed horizontal line), the low-high (dashed lines) and the high-high (dotted lines) procedure of the pilot design controlling the FWER (Figure 3.13 (A)) and the FDR (Figure 3.13 (B)) ($\alpha = 0.05$) for varying c_2 and effect size ratios $k = 3$ and $k = 4$. The solid lines mark the respective maximum over the three procedures for the high-high design. Note that in the unknown variance case we need anyway $n_1 \geq 2$ in order to be able to estimate the variance.

For FWER control (Figure 3.13 (A)) the crossing point between the high-high and the low-high design is at $c_2 = 6.05$ for $k = 4$ and at $c_2 = 4.82$ for $k = 3$. The corresponding asymptotic optimal first stage sample sizes n_1 for the high-high design at those crossing points are 2.29 for $k = 4$ and 2.75 for $k = 3$. The crossing point between the low-high design and the low-low design is $c_2 = 9.02$ for $k = 4$ and at $c_2 = 6.31$ for $k = 3$. Note that the power of the low-low design controlling the FWER is always 0.555 (see Table 3.5).

For FDR control the results are similar (see Figure 3.13 (B)). The crossing point between the high-high and the low-high design is at $c_2 = 6.39$ for $k = 4$ and at $c_2 = 5.06$ for $k = 3$. The corresponding n_1 for the high-high design at the crossing points are 2.14 for $k = 4$ and 2.58 for $k = 3$. The crossing point between the low-high design and the low-low design is $c_2 = 9.25$ for $k = 4$ and at $c_2 = 6.44$ for $k = 3$. The power of the low-low design controlling the FDR is 0.608 for all values of c_2 (see Table 3.5).

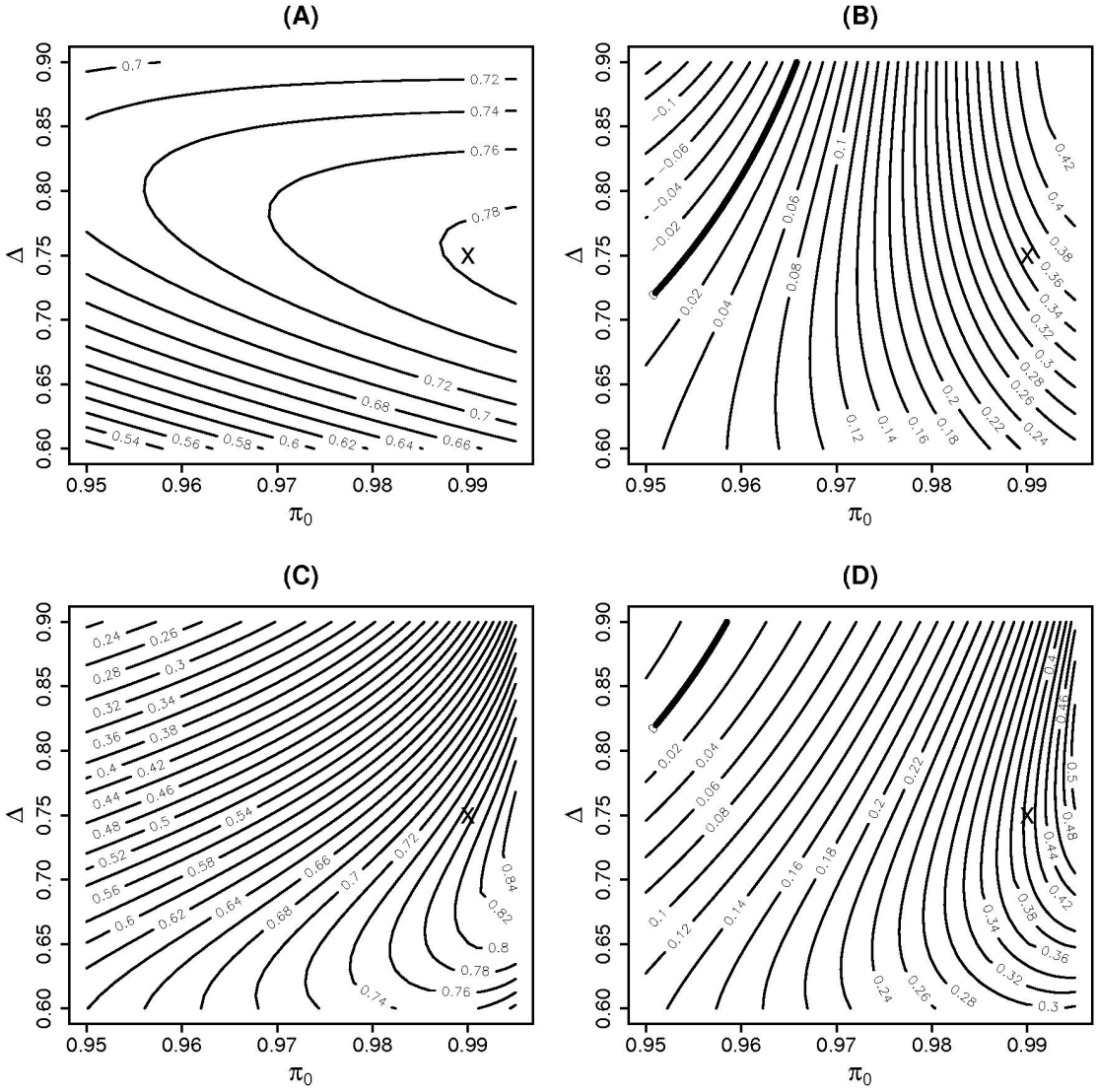

Figure 3.12: Unknown variance case: Contour plots for the difference in power between the single-stage and the pilot design as a function of the true π_0 and Δ for controlling the FWER (first row) or the FDR (second row) for $c_2 = 1$ (first column) and $c_2 = 15$ (second column). Positive values indicate superiority of the two-stage design. Asymptotically optimal two-stage designs were planned for $\pi_0 = 0.99$ and $\Delta = 0.75$ (marked as cross, confer Table 3.5).
$C = 20000$, $m_1 = 1000$, FWER and FDR both $\alpha = 0.05$.

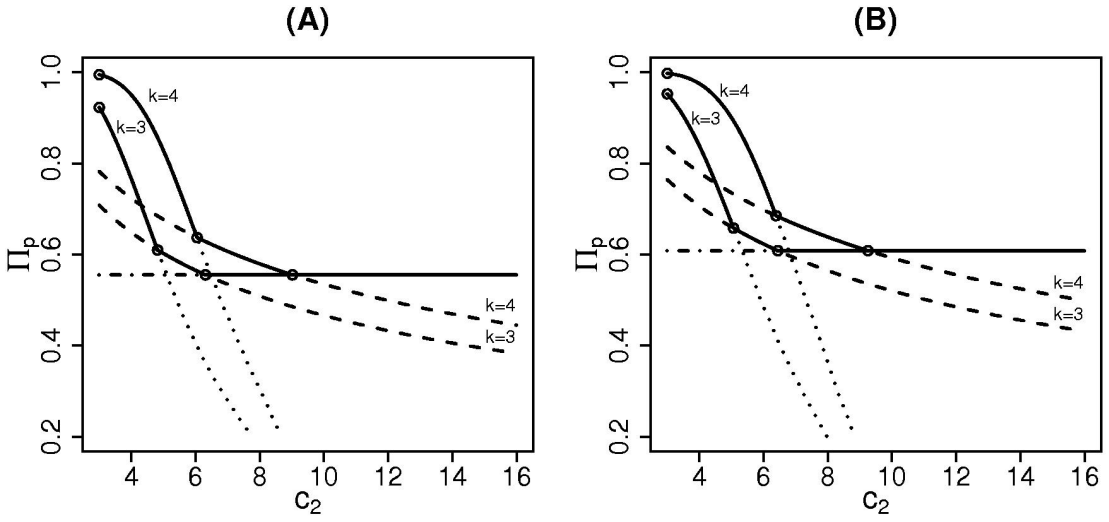

Figure 3.13: Unknown variance case: Asymptotically optimal power for the unknown variance case of the low-low (dotdashed horizontal line), the low-high (dashed lines) and the high-high (dotted lines) procedure of the pilot design controlling the FWER (A) and the FDR (B) for varying c_2 and effect size ratios $k = 3$ and $k = 4$. The solid lines mark the respective maximum over the three procedures for the high-high design. $C = 20000$, $m_1 = 1000$, $\pi_0 = 0.99$, $\Delta = 0.5$, FWER and FDR both $\alpha = 0.05$.

3.9.2 Correlated hypotheses

To investigate the impact of correlation we assume an autoregressive correlation structure among the hypotheses. The correlation between hypothesis i and j is given by $\rho^{|i-j|}$ for some $\rho \in (0, 1)$. The alternative hypotheses are randomly distributed among the $m_1 = 1000$ hypotheses. The effect size for the alternatives is assumed to be $\Delta = 0.75$ and the proportion of true null hypotheses $\pi_0 = 0.99$. We considered the situations of $c_2 = 1, 5$ and 15 assuming $\rho = 0.2, 0.6$ and 0.9 for the pilot design controlling the FWER and FDR by simulation. The total costs C were again set to 20000. Asymptotically optimal parameters γ_1 and r of the uncorrelated case are used (compare Tables 3.1 and 3.2) for the simulations (100000 simulated samples).

The results for FWER control are shown in Table 3.6. The selection boundary increases with increasing correlation whereas the power decreases with increasing correlation. Since we use the Bonforroni correction, the FWER is controlled despite the correlation structure. It is

Table 3.6: **Simulation for the pilot design controlling the FWER ($\alpha = 0.05$) assuming correlated hypotheses (**100000 **simulation steps).**
Simulation results ($\prod_p^s$, γ_2^s, and α^s) under the constraint of autocorrelated hypotheses for different c_2 and ρ. Asymptotically optimal parameters γ_1 and r of the uncorrelated case were used (compare Tables 3.1 and 3.2).
$\Delta = 0.75$, $k = 1$, $C = 20000$, $m_1 = 1000$, $\pi_0 = 0.99$.

ρ	c_2	$\prod_p^s$ (SD)	γ_2^s (SD)	α^s	m_2^s (SD)
0.2	1	0.923 (0.084)	0.00050 (0.00005)	0.044	100.54 (10.00)
	5	0.753 (0.133)	0.00196 (0.00037)	0.035	26.39 (4.64)
	15	0.574 (0.145)	0.00404 (0.00108)	0.025	13.14 (3.07)
0.6	1	0.921 (0.086)	0.00051 (0.00007)	0.042	100.52 (13.96)
	5	0.747 (0.137)	0.00200 (0.00047)	0.033	26.37 (5.84)
	15	0.571 (0.149)	0.00414 (0.00131)	0.024	13.12 (3.59)
0.9	1	0.909 (0.099)	0.00054 (0.00018)	0.030	100.26 (28.49)
	5	0.728 (0.160)	0.00228 (0.00109)	0.022	26.42 (11.19)
	15	0.556 (0.174)	0.00478 (0.00256)	0.015	13.12 (6.34)

interesting to see that the FWER decreases with increasing correlation and with increasing c_2. With increasing c_2 less hypotheses are selected for the second stage and thus, less true null hypotheses. Hence the FWER level cannot be fully exploited.

Simulation results ($\prod_p^s$, γ_2^s, α^s and m_2^s) under the constraint of autocorrelated hypotheses for the control of the FDR are shown in Table 3.7. As for the FWER control, the power decreases and the selection boundary increases with increasing correlation. It has to be mentioned that for large costs the number m_2 of selected hypotheses may become small, so that the finite sample size modification proposed by Storey et al. (2004) has to be used in order to guarantee control of the FDR. Simulation results using the common estimator ($\hat{\pi}_0$) or the finite sample estimator ($\hat{\pi}_0^*$) are given. Using the finite sample estimator leads to a slight decrease in power. However, for large correlation ($\rho = 0.9$) the average FDR over the simulated samples is larger than the targeted level of $\alpha = 0.05$. For $\rho = 0.6$ the finite sample modification helps to control the FDR. The FDR is increasing with increasing correlation.

Figures 3.14 and 3.15 show boxplots of the actual FDR for the 100000 simulated sam-

Table 3.7: **Simulation for the pilot design controlling the FDR ($\alpha = 0.05$) assuming correlated hypotheses (**100000 **simulation steps).**
Simulation results ($\prod_p^s$, γ_2^s, m_2^s and α^s) under the constraint of autocorrelated hypotheses for different c_2 and ρ. Simulation results using the common estimator ($\hat{\pi}_0$) or the finite sample estimator ($\hat{\pi}_0^*$) are given. Asymptotically optimal parameters γ_1 and r of the uncorrelated case were used (compare Tables 3.1 and 3.2). $\Delta = 0.75$, $k = 1$, $C = 20000$, $m_1 = 1000$, $\pi_0 = 0.99$.

ρ	c_2		$\prod_p^s$ (Sd)	γ_2^s (SD)	α^s (SD)	m_2^s (SD)
0.2	1	$\hat{\pi}_0$	0.941 (0.076)	0.00430 (0.00082)	0.048 (0.067)	127.37 (11.23)
		$\hat{\pi}_0^*$	0.941 (0.076)	0.00423 (0.00079)	0.047 (0.067)	127.35 (11.19)
	5	$\hat{\pi}_0$	0.802 (0.130)	0.01973 (0.01104)	0.052 (0.077)	32.65 (5.25)
		$\hat{\pi}_0^*$	0.800 (0.131)	0.01767 (0.00755)	0.047 (0.073)	32.65 (5.25)
	15	$\hat{\pi}_0$	0.659 (0.158)	0.07111 (0.12176)	0.064 (0.102)	15.74 (3.34)
		$\hat{\pi}_0^*$	0.649 (0.160)	0.04145 (0.03323)	0.046 (0.081)	15.72 (3.34)
0.6	1	$\hat{\pi}_0$	0.939 (0.077)	0.00438 (0.00106)	0.049 (0.072)	127.93 (15.93)
		$\hat{\pi}_0^*$	0.939 (0.077)	0.00429 (0.00102)	0.048 (0.071)	127.40 (15.96)
	5	$\hat{\pi}_0$	0.798 (0.134)	0.02114 (0.01628)	0.054 (0.084)	32.65 (6.75)
		$\hat{\pi}_0^*$	0.796 (0.134)	0.01857 (0.01043)	0.049 (0.079)	32.65 (6.72)
	15	$\hat{\pi}_0$	0.654 (0.164)	0.08635 (0.16059)	0.068 (0.111)	15.76 (3.98)
		$\hat{\pi}_0^*$	0.646 (0.166)	0.04453 (0.03985)	0.047 (0.087)	15.75 (3.96)
0.9	1	$\hat{\pi}_0$	0.930 (0.089)	0.00505 (0.00399)	0.052 (0.110)	127.52 (32.81)
		$\hat{\pi}_0^*$	0.930 (0.089)	0.00054 (0.00312)	0.051 (0.107)	127.32 (32.98)
	5	$\hat{\pi}_0$	0.777 (0.164)	0.05276 (0.13547)	0.074 (0.148)	32.61 (13.16)
		$\hat{\pi}_0^*$	0.774 (0.164)	0.03116 (0.04876)	0.062 (0.131)	32.60 (13.11)
	15	$\hat{\pi}_0$	0.634 (0.197)	0.23403 (0.35788)	0.089 (0.161)	15.73 (7.17)
		$\hat{\pi}_0^*$	0.626 (0.198)	0.07240 (0.07949)	0.057 (0.130)	15.75 (7.22)

ples for the pilot design under the constraint of correlated hypotheses for varying c_2 and ρ using the common and the finite sample estimator.

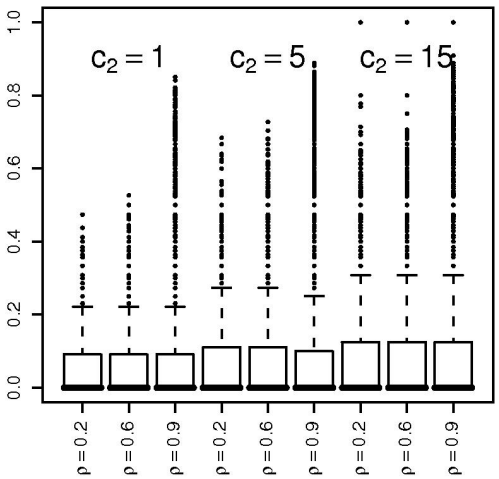

Figure 3.14: Boxplots of the actual FDR for the 100000 simulated samples for the pilot design under the constraint of correlated hypotheses for varying c_2 and ρ using the common estimator for $\hat{\pi}_0$. $\Delta = 0.75$, $k = 1$, $C = 20000$, $m_1 = 1000$, $\pi_0 = 0.99$, targeted $\alpha = 0.05$.

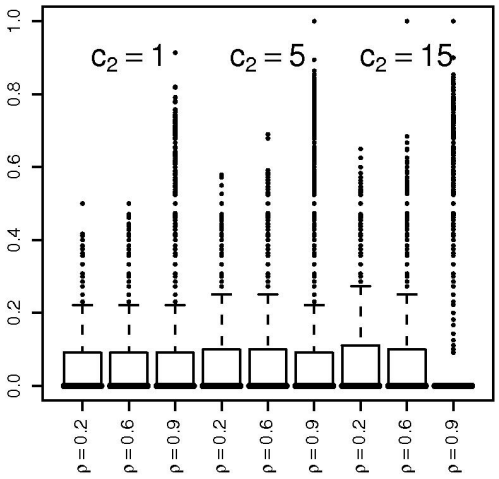

Figure 3.15: Boxplots of the actual FDR for the 100000 simulated samples for the pilot design under the constraint of correlated hypotheses for varying c_2 and ρ using the finite sample estimator $\hat{\pi}_0^*$. $\Delta = 0.75$, $k = 1$, $C = 20000$, $m_1 = 1000$, $\pi_0 = 0.99$, targeted $\alpha = 0.05$.

3.9.3 Integer stage-wise sample sizes

In the examples investigated, we considered r as continuous variable and thus the optimal sample sizes per stage (n_1 and n_2) in general will be non-integer. Since in reality we need integer sample sizes, we performed simulations for the pilot design using the optimal r and γ_1 from the non-integer optimization. The situations from Tables 3.1 and 3.2 for $\Delta = 0.75$ are investigated by simulation.

In order to achieve constant costs we first rounded the non-integer first stage costs rC downwards and the second stage costs $(1-r)C$ upwards. For the first stage sample sizes we used $\lfloor \lfloor rC \rfloor / m_1 \rfloor + 1$ for $\lfloor rC \rfloor$ mod m_1 randomly chosen hypotheses and $\lfloor \lfloor rC \rfloor / m_1 \rfloor$ for the rest. For the second stage sample sizes we performed similar: we used $\lfloor \lceil (1-r)C \rceil / (m_2 c_2) \rfloor + 1$ for $\lceil (1-r)C \rceil$ mod $m_2 c_2$ randomly chosen hypotheses and $\lfloor \lceil (1-r)C \rceil / (m_2 c_2) \rfloor$ for the rest. $\lfloor x \rfloor$ denotes the largest integer not exceeding x and $\lceil x \rceil$ the smallest integer exceeding x. Note that the simulated C may be slightly smaller than the targeted $C = 20000$.

Using designs with stage-wise integer sample sizes does not noticeable decrease the power as compared to the optimal non-integer designs (see Tables 3.1 and 3.2). Table 3.8 gives the results for the simulations under the constraint of integer stage-wise sample sizes for different c_2 for the control of the FWER or FDR. Results for using the finite sample size modification to guarantee control of the FDR are given. Using this modification leads to a slight decrease in power. Table 3.8 shows simulation results for the control of the FDR using the common estimator ($\hat{\pi}_0$) and using the finite sample estimator ($\hat{\pi}_0^*$). Using the common estimator for $c_2 = 5, 15$ the mean FDR (over the simulated samples) is larger than the targeted FDR of 0.05. Using the finite sample modification in average the FDR is controlled. Figure 3.16 shows the Boxplots of the actual FDR for the 100000 simulated samples for the pilot design under the constraint of integer stage-wise sample sizes for varying c_2 using the common (first 3 boxplots) and the finite sample estimator (last 3 boxplots).

Table 3.8: **Simulation results for the pilot design controlling the FWER or FDR ($\alpha = 0.05$) under the constraint of integer stage-wise sample sizes (100000 simulation steps).**
Simulation results ($\prod_p^s$, γ_2^s, m_2^s and α^s) under the constraint of integer stage-wise sample sizes for different c_2. For the FDR control, simulation results using the common estimator for π_0 ($\hat{\pi}_0$) and using the finite sample modification ($\hat{\pi}_0^*$) are given. Asymptotically optimal parameters γ_1 and r using non-integer sample sizes were used (Tables 3.1 and 3.2).
$\Delta = 0.75$, $k = 1$, $C = 20000$, $m_1 = 1000$, $\pi_0 = 0.99$.

		c_2	$\prod_p^s$ (Std)	γ_2^s (Std)	α^s	m_2 (Std)
FWER		1	0.924 (0.084)	0.00050 (0.00005)	0.044	100.54 (9.15)
		5	0.753 (0.133)	0.00195 (0.00035)	0.034	26.38 (4.44)
		15	0.574 (0.146)	0.00403 (0.00107)	0.025	13.14 (3.02)
FDR	$\hat{\pi}_0$	1	0.941 (0.075)	0.00433 (0.00076)	0.047 (0.067)	127.41 (10.20)
	$\hat{\pi}_0^*$		0.941 (0.075)	0.00422 (0.00076)	0.047 (0.077)	127.44 (10.24)
	$\hat{\pi}_0$	5	0.802 (0.129)	0.01970 (0.01105)	0.052 (0.103)	32.65 (5.02)
	$\hat{\pi}_0^*$		0.800 (0.130)	0.01758 (0.00750)	0.047 (0.066)	32.66 (5.01)
	$\hat{\pi}_0$	15	0.660 (0.157)	0.07028 (0.11881)	0.065 (0.072)	17.73 (3.25)
	$\hat{\pi}_0^*$		0.649 (0.160)	0.04098 (0.03244)	0.045 (0.081)	15.73 (3.26)

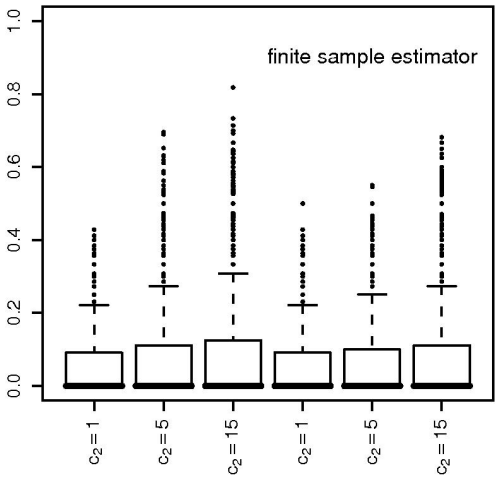

Figure 3.16: Boxplots of the actual FDR for the 100000 simulated samples for the pilot design under the constraint of integer stage-wise sample sizes for varying c_2 using the common (first 3 boxplots) and the finite sample estimator (last 3 boxplots) for π_0. $\Delta = 0.75$, $k = 1$, $C = 20000$, $m_1 = 1000$, $\pi_0 = 0.99$, targeted $\alpha = 0.05$.

3.10 Discussion

We have investigated two-stage designs in the situation that large numbers of null hypotheses are tested and only a small proportion of them are expected to be wrong. Moreover it was assumed that there are constraints on total costs of the experiment. The first stage is used for screening out promising hypotheses which are then investigated further at the second stage. We focussed on an important scenario in practice assuming that costs per measurement differ between stages: On the one hand extra costs may arise when the same measurements have to be designed for a subset of hypotheses selected in an interim analysis and investigated at the second stage. On the other hand the investigator from the very beginning may have the choice between a low-cost method and a high-cost method, which hopefully is more efficient in terms of the effect size under the alternatives. Given a large number of candidate hypotheses we derived asymptotically optimal designs in terms of power using the simplifying assumptions of common alternatives (either controlling the FWER or the FDR).

We would like to summarize the results in the following way:
Two-stage screening designs are a very powerful tool even if we deal with equal effect sizes at the second stage but the costs for designing the measurements for the selected hypotheses at the second stage are fairly high. Only severe design misspecifications in the planning phase may lead to a noticeable loss of power such that the single-stage design may become superior in power. With regard to the impact of design misspecifications in the proportion of true alternatives it seems to be preferable not to assume too small proportions in the planning phase. Integrated designs which use data from both stages for the final test decisions are more robust against design misspecifications.

If two different methods are available, depending on the ratios between costs and effect sizes it is preferable to run two-stage designs which apply either the low-cost or the high-cost method at both stages. Designs starting with the low-cost method and switching to the more expensive method in the interim analysis may only be advisable if there is lack of resources

so that first stage sample size for the high-cost method would be too small. However, it has to be kept in mind that the best design depends on the relationship of the effect size and the cost ratios. Hence in case of a effect size misspecifications in the planning phase the low-high method may actually be more powerful than the low-low or the high-high strategy. However it seems natural to apply a design which is preferable under the parametric constellation considered in the planning phase. In the integrated design the optimal way of combining more over data from both stages arising from different measurement methods depends on the effect size ratio between stages, which introduces a further complication for appropriately designing such experiments applying different methods.

With respect to deviations from the underlying assumption we calculated optimal designs for the unknown variance case. The results for r and γ_1 are very close to those of the known variance case. The optimal power decreases as compared to the known variance case. However, using the optimal parameters for the known variance case in the situation of unknown variances leads to virtually the same performance as using the optimal parameters from the unknown variance case. Concerning the second scenario, the decision which of the procedures (low-low, high-high or low-high) is preferable is more difficult because no common crossing point in costs as a function of c_2 between the three procedures exists. However, the region where the low-high procedure is preferable still remains small.

To investigate the impact of correlation we assumed an autoregressive correlation structure among the hypotheses. The alternative hypotheses were randomly distributed among the hypotheses. The impact of correlation was small like in the case of constant costs in Zehetmayer et al. (2005). For the two-sided situation we refer to their proposal to test a set of $2m_1$ one-sided hypotheses.

4 Prognostic scores based on gene expression or proteomic data

4.1 Introduction

When a large number of markers have to be investigated, we usually can not trust that a few of these markers show big effects. Instead we may hope that there is a combination of several markers which, e.g., allow a prediction of the response of an individual patient to a particular therapy. The task of selecting useful markers with rather moderate effects from a (very) large number of candidates and estimating suitable scores to be used for the prediction in future patients is a formidable exercise. Moreover, due to limited resources generally small sample sizes per marker are available which makes the problem even less tractable. It seems that for a lot of medical research reported in this field there has been insufficient awareness of the statistical properties of the resulting prognostic scores. Ntzani and Ioannidis (2003) showed for prediction of cancer outcome that the constructed scores are poorly performing in external validation samples. The better performance in a few small studies may also be due to publication bias.

Subset selection procedures (e.g. Shao (1993), Miller (2002)) are widely used in this field. However, there is a general problem of how to quantify the probability for falsely selecting ineffective parameters. There have been proposals of estimating the positive false discovery rate in case that a nonzero model has been selected (Li and Hui (2007)). It is known that model selection by multiple testing of individual model parameters under fairly general

conditions asymptotically is a consistent selection procedure: for increasing sample size the critical boundary for the individual test statistics (the parameter estimate divided by its standard error) has to approach infinity at a smaller order than the inverse of the standard error (see Bauer, Pötscher and Hackl (1988)). Asymptotic relationships between model selection procedures and multiple tests controlling the False Discovery Rate (FDR, see Section 2.1) have been shown (Abramovich, Benjamini, Donoho and Johnstone (2006)). However these are general results which do not help how to choose the selection boundaries in a specific sample.

In the following this problem of constructing prognostic scores is discussed. The multiple testing type methods of selecting markers and constructing a linear prognostic score are considered in terms of the statistical properties of the resulting receiver operating characteristic (ROC-curve) in future patients, e.g., in terms of how well they can predict the outcome of a medical therapy in future patients. Various measures have been proposed to capture discrimination, but the area under the ROC-curve (AUC) is the most popular quantity (see e.g. Pencina et al. (2008)). It is a measure of the predictive ability of a score if the score is used for different thresholds with varying values of sensitivity and specificity. Note that the AUC can be interpreted as the probability that in a randomly selected (according to a uniform distribution) pair of responders and non-responders, the score value of the non-responder is smaller than the score value of the responder.

For different scenarios we assume that in the population a set of markers exists which, if known, would lead to a prediction score with a ROC-curve crossing a benchmark point with pre-fixed benchmark values for sensitivity and specificity.

We discuss three selection methods:

- multiple testing of individual hypotheses controlling the FDR,
- using a stepwise selection procedure by testing individual parameters and
- selecting the k "best" markers in an unprotected way.

Simulations for independent normal distributions with known variance will be presented for a varying number of tested markers, varying sample sizes and varying proportion of effective markers with equal effect sizes. To achieve a standardization, the effect sizes are chosen such that the optimal linear prediction of future patients, if known, would lead to an ROC-curve crossing through the benchmark point where sensitivity and specificity are equal 0.9, thus the theoretically achievable area under the ROC-curve (AUC) is 0.965. Finally the situation is considered that cross validation is used to determine decision boundaries for the test based selection procedure optimal with regard to the AUC, thus achieving some information on the extent of false positive decisions.

Section 4.2 describes the investigated problem. In Section 4.3 the three selection methods are discussed and in Section 4.4 we explain the prediction scores based on these selection methods. The results of the performed simulations can be seen in Section 4.6. Using cross validation to determine decision boundaries is discussed in Section 4.7. In Section 4.8 we investigate simulations assuming a smaller best theoretically achievable AUC, further on denoted by AUC_*, of 0.8. A short discussion of all results is given in Section 4.9.

4.2 The problem

Consider the following simple scenario: we want to search for predictors of a specific clinical outcome (e.g. effectiveness of certain chemotherapy) among a large set of m markers. Independent samples of patients responding (n_r) and non-responding (n_{nr}) to a specific therapy are available. Based on the marker values in responding and non-responding patients effective markers have to be selected and a score to predict response to therapy of future patients has to be estimated. This type of situation may be encountered in many other areas of research based on clinical data looked at together with gene expression or proteomic data.

Assume that the marker levels follow normal distributions with common known variance $\sigma^2 = 1$ with means $\mu_{r,i}$, $i = 1, ..., m$, for responders and means $\mu_{nr,i}$, $i = 1, ..., m$, for non-

responders.

In the following we will first investigate selection procedures in a single training sample which apply at least also for individual steps in model selection procedures, e.g. with cross validation (considered later).

4.3 Selection of markers

We investigate three methods for the selection of promising predictors for a clinical outcome of a future patient.

4.3.1 The protected approach based on a multiple test controlling the false discovery rate

For the selection of prognostic markers for the prediction score we test the following set of one sided null hypotheses:

$$H_{0i} : \mu_{r,i} - \mu_{nr,i} = 0 \text{ against } H_{1i} : \mu_{r,i} - \mu_{nr,i} > 0, \text{ for } i = 1, \ldots, m.$$

The standardized mean differences between responders and non-responders

$$z_i = (\bar{x}_{r,i} - \bar{x}_{nr,i})/\sqrt{2/n}, \qquad i = 1, \ldots, m$$

are calculated, where for simplicity we assume equal sample sizes per marker and group $n = n_{r,i} = n_{nr,i}$ for $i = 1, \ldots, m$. Test decisions are based on the one sided p-values $p_i = 1 - \Phi(z_i)$, where Φ denotes the cumulative distribution function of the standard normal distribution. To adjust for multiplicity, i.e. not to include too many nuisance markers without any predictive ability in the score, we use Storey's approach (see Storey (2002)) to control the FDR discussed in Section 2.3.2 (see Formula (2.6)) where the critical boundary γ is estimated from the sample. Note that this method adapts to the estimated proportion of true null hypotheses. The markers whose p-values fall below the critical boundary γ are selected to build a score in order to predict whether a future patient will respond or not respond to

the treatment.

Note that for the two-sided p-values ($p_i = 1 - \Phi(|z_i|)$) the same procedure applying the critical boundary $\frac{\gamma}{2}$ leads to the same results as the one-sided test under the global null hypotheses. If under the alternative we assume constant effect size among the effective markers (as in the following) the two-sided tests lead to the same results if the effect size is increased by $(c_{1-\frac{\gamma}{2}} - c_{1-\gamma})/\sqrt{\frac{n}{2}}$ where $c_{1-\gamma}$ denotes the $(1-\gamma)$-quantile of the standard normal distribution (ignoring directional errors under the alternative).

4.3.2 Selection using stepwise forward logistic regression

Whereas the first approach is only based on individual selection criteria for the markers, here we use a multiple logistic regression approach with stepwise forward selection to assess the contribution of the individual markers to predict response to therapy in the training sample. We again use a fixed threshold γ for the p-values, this time calculated from the final model evolving in the multivariate logistic regression. The selection is done in such a way that selected markers can again be removed from the model when their p-values in the aggregated model fall above the threshold γ. The stepwise procedure ends with a final model when further markers fail to meet the selection criterion.

4.3.3 The optimistic approach by selecting the k best markers

Alternatively to the FDR-controlled approach, believing in the effect of some of the markers, simply the best k genes are selected for prediction. I.e., the genes with the k smallest individual one-sided p-values are used to predict the clinical outcome of an independent future patient.

4.4 Prediction of the clinical outcome

After selecting k hypotheses by one of the three methods described we simply use a linear combination of the selected markers as a prognostic score. Exploiting the assumptions of independence across the markers and known variances, we simply use the linear score which would be the Bayes solution for a given (unselected) set of k prognostic variables which follow independent and identical normal distributions with unknown means and known variance $\sigma^2 = 1$ (e.g. see Anderson (2003)):

$$\hat{f}(\mathbf{x}) = (\hat{\boldsymbol{\mu}}_r - \hat{\boldsymbol{\mu}}_{nr})^T \mathbf{x} = \sum_{i=1}^{k} (\bar{x}_{r,i} - \bar{x}_{nr,i}) x_i. \tag{4.1}$$

Here to simplify notation we have rearranged the markers so that without loss of generality $\hat{\boldsymbol{\mu}}_r^T = (\bar{x}_{r,1}, ..., \bar{x}_{r,k})$ and $\hat{\boldsymbol{\mu}}_{nr}^T = (\bar{x}_{nr,1}, ..., \bar{x}_{nr,k})$ are the means of the k selected markers in the training samples of the patients responding and not responding to therapy respectively and $\mathbf{x} = (x_1, ..., x_k)$ are the values of the corresponding markers in a future patient (bold symbols give vectors). If $\hat{f}(\mathbf{x}) > c^*$ we predict a response, otherwise a non-response. The predictive ability of such a score is assessed by the ROC curve resulting from varying threshold values for the score, where sensitivity (for response) is plotted against (1-specificity) as a function of c^*. Since we are interested in ROC curves the following results are invariant to any strictly monotonic transformation of this score. Note that this discriminant analysis type of predictors assuming diagonal covariance matrix is often also applied as an approximation if correlations between the markers can not be excluded and under the known variance assumption corresponds to the naive Bayes predictor.

In case of forward selection we simply use the estimated linear predictor for the log odds from the final model in the stepwise multiple logistic regression, which is of the same form as (4.1) but uses the parameter estimates from the model instead of the difference in the sample means of the selected markers. The score (4.1) is also used for the k selected best markers in the optimistic approach.

Given the estimates from the training samples the prognostic score (4.1) follows a normal

distribution:

$$N[\mu_f, \sigma_f^2] = N[(\hat{\boldsymbol{\mu}}_r - \hat{\boldsymbol{\mu}}_{nr})^T \boldsymbol{\mu}, (\hat{\boldsymbol{\mu}}_r - \hat{\boldsymbol{\mu}}_{nr})^T (\hat{\boldsymbol{\mu}}_r - \hat{\boldsymbol{\mu}}_{nr})] \tag{4.2}$$

where $\boldsymbol{\mu}^T = (\mu_1, \ldots, \mu_k)$ is the true mean vector of the k selected markers in an independent future patient. Fixing $\boldsymbol{\mu}$ it is easy to get the ROC-curve for future independent populations of responders and non-responders by calculating

$$Sensitivity = v = 1 - \Phi_{\mu_f, \sigma_f^2}(c^*) = 1 - \Phi(\frac{c^* - \mu_f}{\sigma_f})$$

and

$$Specificity = 1 - w = \Phi_{0, \sigma_f^2}(c^*) = \Phi(\frac{c^*}{\sigma_f})$$

where Φ_{μ,σ^2} denotes the cumulative distribution function of the normal distribution with mean value μ and variance σ^2. Thus, the AUC can be calculated as:

$$AUC = \int_0^1 (1 - \Phi(c_{1-w} - \frac{\mu_f}{\sigma_f}))dw. \tag{4.3}$$

Note that given the training samples the vector $\boldsymbol{\mu}$ depends on the selection procedure and may also contain means from selected ineffective markers not contributing to prognosis.

4.5 Minimal effect size Δ

To simplify the problem we assume that the effective markers have a common mean $\mu_{r,i} = \mu_r$, $i \in E$ in the responding patients and a common mean $\mu_{nr,i} = \mu_{nr}$, $i \in E$ in the non-responding patients, E denoting the index set of the $m_e = |E|$ effective markers among the m candidates. For the non-effective markers without loss of generality the mean difference between responders and non-responders is assumed to be zero, $\mu_{r,j} - \mu_{nr,j} = 0$, $j \in (1, 2, .., m) \setminus E$. Hence for the effective markers a common effect size $\mu_r - \mu_{nr} = \Delta$ is assumed to hold. In this scenario a linear function of the values of the effective markers (with equal weights) would be the optimal score for prediction.

To get a benchmark let us assume that this optimal linear score built from the m_e effective markers among the set of m markers is known. We now will ask, depending on the

number m_e of effective markers, what minimal common effect size Δ is required to achieve a ROC-curve crossing through the point where both sensitivity (v) and specificity ($1-w$) have a certain pre-specified values, e.g., $v = 1-w = 0.9$? Clearly we get the best prognostic score if all m_e effective markers and no non-effective markers are selected, and the true effect size Δ is known:

$$f(\mathbf{x}) = (\hat{\boldsymbol{\mu}}_r - \hat{\boldsymbol{\mu}}_{nr})^T\mathbf{x} = \boldsymbol{\Delta}^T\mathbf{x} = \Delta\Sigma_{i=1}^{m_e}x_i,$$

which follows a normal distribution with:

$$N[\mu_f, \sigma_f^2] = N[\boldsymbol{\Delta}^T\boldsymbol{\mu}, \boldsymbol{\Delta}^T\boldsymbol{\Delta}] = N[m_e\Delta^2, m_e\Delta^2]$$

Hence the sensitivity for the theoretically best score for a future patient can be easily calculated as follows:

$$v = 1 - \Phi(c_{1-w} - \frac{m_e\Delta^2}{\sqrt{m_e\Delta^2}}) = 1 - \Phi(c_{1-w} - \sqrt{m_e}\Delta)$$

and thus the effect size required to cross the point $(v, 1-w)$ can be calculated as:

$$\Delta = \frac{c_{1-w} - c_{1-v}}{\sqrt{m_e}} \tag{4.4}$$

Figure 4.1 shows the minimal effect size Δ depending on the number of effective markers m_e if the ROC-curve crosses the point $v = 1-w = 0.9$. Two examples are marked which will be considered more closely in the simulation studies. For $m_e = 60$ an effect size of $\Delta = 0.33$ is required to achieve such an ROC-curve. For $m_e = 10$ an effect size of $\Delta = 0.81$ is needed to get a ROC-curve with such a property. Note that if there is only a single effective marker an effect size of $\Delta = 2.56$ is required. This demonstrates the crucial problem for gene expression studies. If 60 efficient markers work together they may show a large common effect even if there are only marginal individual effect sizes. Thus, the process of selection of such markers with only marginal effects among a large number of candidates in relatively small samples will be a formidable task. However, in case of a single or few efficient markers the effect size to achieve good prognostic properties has to be pretty large, so that already small samples may be sufficient to select those very influential markers.

Figure 4.2 shows the minimal effect size Δ required for the ROC-curve to cross through the points $v = 1 - w = 0.9$, 0.8 and 0.7 as a function of m_e. Clearly, to obtain a smaller sensitivity and specificity smaller effect sizes are required. E.g. to achieve a ROC-curve crossing through the point $v = 1 - w = 0.8$ an effect size of $\Delta = 0.53$ is needed if $m_e = 10$ and a very small Δ of 0.22 if $m_e = 60$.

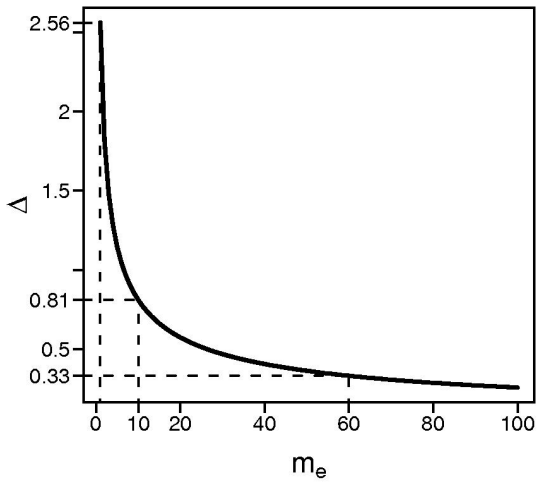

Figure 4.1: Dependence of Δ on m_e: Minimal effect size required for the ROC to cross through the point where sensitivity and specificity are equal to 0.9 for a varying number of effective markers m_e.

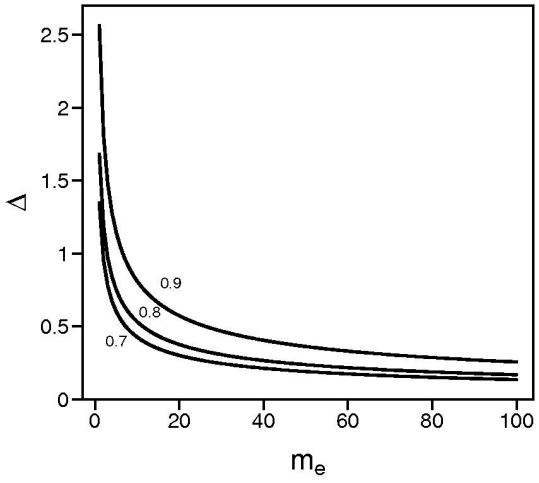

Figure 4.2: Dependence of Δ on m_e: Minimal effect size required for the ROC to cross through the points where sensitivity and specificity are equal to 0.9, 0.8 and 0.7 as a function of the number of effective markers m_e.

4.6 Simulations results

We will now investigate the three different selection procedures for constructing a linear score discussed in Section 4.4 by simulation, assuming that two samples of patients responding to a particular treatment and of patients not responding to the treatment are available. We fix the sample sizes as $n_r = n_{nr} = n = 50$, 100 and 500 per group and the candidate marker measurements ($m = 1000$ and 6000) are assumed to follow independent normal distributions with common known variance $\sigma^2 = 1$. For $m\pi_0$ markers the means are equal (the null hypothesis is true) and for the remaining $m_e = m(1-\pi_0)$ markers the same alternative $\mu_{r,i} - \mu_{nr,i} = \Delta$, $i \in E$ holds. Hence, the true mean differences between responders and non-responders are Δ for the k_e ($k_e \leq k$) selected effective markers. For the $k_0 = k - k_e$ selected ineffective markers the mean differences are 0. Looking at the mean μ_f and the variance σ_f^2 (see formula (4.2)) of the prediction score (4.1) it is easy to see that the higher the proportion of ineffective markers among the selected ones the smaller the mean distance between the two homoscedastic normal distributions of future responders and non responders, and hence the worse the prognosis.

We vary the number of effective markers to be $m_e = 10$, 60 or 0. Different FDR values

for using the protected procedure and different numbers k of "best" markers for using the unprotected procedure are investigated. Also different boundaries γ when using the forward logistic regression as selection method are investigated.

4.6.1 Variable selection using the protected approach

As discussed in Section 4.5, the effect size Δ is triggered by forcing the optimal ROC-curve through the benchmark point $v = 1 - w = 0.9$, thus for $m_e = 10$, $\Delta = 0.81$ and for $m_e = 60$, $\Delta = 0.33$.

m_e=10, m=1000

n=50

The sample size per group is first fixed to $n = 50$. Figures 4.3 show simulated ROC-Curves (grey curves) for some FDR selection criteria assuming $m_e = 10$ effective markers (alternatives) among $m = 1000$ hypotheses. The "average" ROC is calculated from the simulated samples, conditional that at least one marker has been identified by the selection procedure (dashed curve). The quantity $\hat{p}_s$ is the proportion of samples among all simulation steps where at least one marker has been identified for future prediction. The theoretically achievable ROC-curve through the benchmark point $(0.9, 0.9)$ is shown by the solid curve.

To compare the different FDR selection criteria we calculated the area under the ROC-curve (AUC) for an independent future patient (see formula (4.3)) for all simulated samples. Figure 4.4 (A) shows boxplots of the AUC-values of the simulated samples (10000 repetitions) for different FDR values chosen a priori for the selection of markers for future prediction, conditional that at least one marker has been identified for future prediction. Note that in the simulations for this scenario $\hat{p}_s = 1$ if the FDR is larger than 0.01. When controlling a FDR of 0.00005, $\hat{p}_s = 0.67$. The best performance, i.e. the largest average AUC among the grid of investigated FDR values, further on denoted by AUC_*^{sim}, occurs if selection is performed con-

trolling a FDR of 0.15 and 0.20. For these two investigated values $AUC_*^{sim} = 0.941$ which is slightly smaller than the theoretically best achievable AUC_* of 0.965 of the ROC-curve crossing through the benchmark point $v = 1 - w = (0.9, 0.9)$ (see the dotdashed horizontal line in the Figures). If a larger FDR is chosen, more ineffective markers are tolerated in the score but also more effective marker are selected so that the score still performs well. Fixing FDR= 0.15 in average 1.7 true null hypotheses and 8.6 alternatives (effective markers) are selected for future prediction. For FDR= 0.20 in average 2.4 true null hypotheses and 8.9 alternatives are selected. This leads to a similar performance in terms of AUC values for future prediction as for FDR= 0.15. However, if a too large FDR is chosen for selection, too many ineffective markers are added so that the score gets worse, e.g. choosing the FDR= 0.95 the average "future" AUC becomes much smaller (0.827) than AUC_*. Although in average all 10 alternatives are then selected, the prediction score additionally includes in average 374.8 true null hypotheses.

n=100

When increasing the sample size per group to $n = 100$ clearly the AUC values of the selected scores also increase. Figure 4.4 (B) shows the situation assuming $m_e = 10$ effective markers among $m = 1000$ tested hypotheses for a sample size of $n = 100$ per group. The FDR with the largest average AUC for future prediction among the investigated values (AUC_*^{sim}) is smaller than in the situation of $n = 50$ per group. Selection controlling a FDR of 0.01 and 0.05 leads to virtually the same performance with an AUC_*^{sim} of 0.961 which is again only slightly smaller than $AUC_* = 0.965$. In average 9.9 alternatives but only 0.6 true null hypotheses are selected for a future prediction score when choosing FDR= 0.05. $\hat{p}_s = 1$ for all investigated FDR selection criteria fixing the sample size to $n = 100$ per group. Figure 4.4 (B) shows that the resulting scores perform well for a wide range of FDR values. This may be due to the better estimate of the true mean values ($\mu_{r,i}$ and $\mu_{nr,i}$ for $i = 1, ..., m$) and thus to a better estimate of the weights of the selected markers in the score function when the sample size increases. Therefore selection of one true null hypothesis (ineffective marker) has a smaller impact on the overall performance than one selected alternative (effective marker),

since its weight in the score function is smaller, although its p-value is still significant for larger FDR values.

n=500

Figure 4.4 (C) shows the situation assuming a sample size of $n = 500$ per group. It seems that for $n = 500$ it does not really mind which FDR is chosen as selection criteria, the resulting score always performs good. Because of the large sample size, the p-values of the alternatives are nearly 0 and thus mostly all alternatives are selected for future prediction. As mentioned before, there is a better estimate of the mean values of the two groups and thus, the weights of selected true null hypotheses are nearly null, although their p-values are significant for larger FDR values. This also interprets the average AUC of 0.945 for selection using a FDR of 0.99. $AUC_*^{sim} = 0.965$ occurs when choosing smaller selection criteria (up to FDR= 0.01) and is equal to AUC_*. Always all 10 effective markers but in average only up to 0.1 ineffective markers (for FDR= 0.01) are selected for the prediction score.

m_e=60, m=1000

n=50

Let us now assume $m_e = 60$ alternatives among the $m = 1000$ tested hypotheses. The sample size per group is again first fixed to $n = 50$. The effect size Δ to achieve the theoretical benchmark ROC-curve now is 0.33. Figures 4.5 again show simulated ROC-Curves (grey curves) as well as the average curve (dashed line) and the theoretical benchmark ROC (solid line) for some FDR selection criteria and Figure 4.6 (A) shows the boxplots of the AUC-values of the simulated samples for the different FDR selection criteria, conditional that at least one marker was selected for future prediction. In this scenario $\hat{p}_s$ in the simulated samples approaches 1 only if the FDR is larger than 0.60.

Again it is better to tolerate more ineffective markers and thus find more effective markers,

however in this scenario it would be superior to chose a very large FDR as selection criterion. Selection controlling a FDR of 0.8 and 0.85 leads to prediction scores with $AUC_*^{sim} = 0.813$, whereas selection controlling a FDR of e.g. 0.2 leads to scores with an average AUC of only 0.691, which indicates a poor performance. Note also that for FDR= 0.2, $\hat{p}_s = 0.918$, that means that in 8.2% of the simulated samples no prediction score is selected from the data. When controlling a FDR of 0.05 in only 58.2% of the simulated samples a prediction score is selected from the data. Choosing FDR= 0.85 in average includes 51.6 alternatives and 325.4 true null hypotheses in the score for future prediction. For FDR= 0.20 in average 2.0 true null hypotheses but only 6.7 out of the $m_e = 60$ alternatives are selected. However, $AUC_*^{sim} = 0.813$ is still much smaller than $AUC_* = 0.965$. It is interesting to see that selection controlling a FDR of 0.95 results in a performance with an average AUC of 0.777 which is still good as compared to small FDR values. However, it remains the question whether a score built with in average more than 800 markers is practical, although the future performance is larger as compared to smaller FDR values.

n=100

Again increasing the sample size per group to $n = 100$ increases the AUC values of the selected scores. Figure 4.6 (B) shows the situation assuming $m_e = 60$ effective markers among $m = 1000$ tested hypotheses fixing the sample size to $n = 100$ per group. Again the corresponding FDR where AUC_*^{sim} occurs is smaller than in the situation of $n = 50$ per group. AUC_*^{sim} is 0.887 for selection controlling a FDR between 0.45 and 0.55. Choosing FDR= 0.55 leads to a score including on average 47.0 effective and 59.8 ineffective markers. For a FDR above 0.10 the simulations show $\hat{p}_s = 1$. Clearly with an increasing sample size also $\hat{p}_s$ increases. Again this may be due to the better estimate of the true mean values $\mu_{r,i}$ and $\mu_{nr,i}$ $(i = 1, ..., m)$.

n=500

Figure 4.6 (C) shows the situation assuming a sample size of $n = 500$ per group. Again

a good performance in terms of future AUC values can be seen over all investigated FDR values. When selection is based on a FDR of 0.99, the average AUC is 0.925. The largest average AUC for prediction of a future patient among the investigated FDR values (AUC_*^{sim}) of 0.961 occurs when choosing a FDR of 0.01 and 0.05. Fixing FDR= 0.05 in average 59.6 out of the 60 alternatives and 3.2 true null hypotheses are selected for future prediction. Again, because of the large sample size, $\hat{p}_s$ is 1 for all investigated FDR values.

For more detailed information see Tables E.8 to E.13 in the appendix.

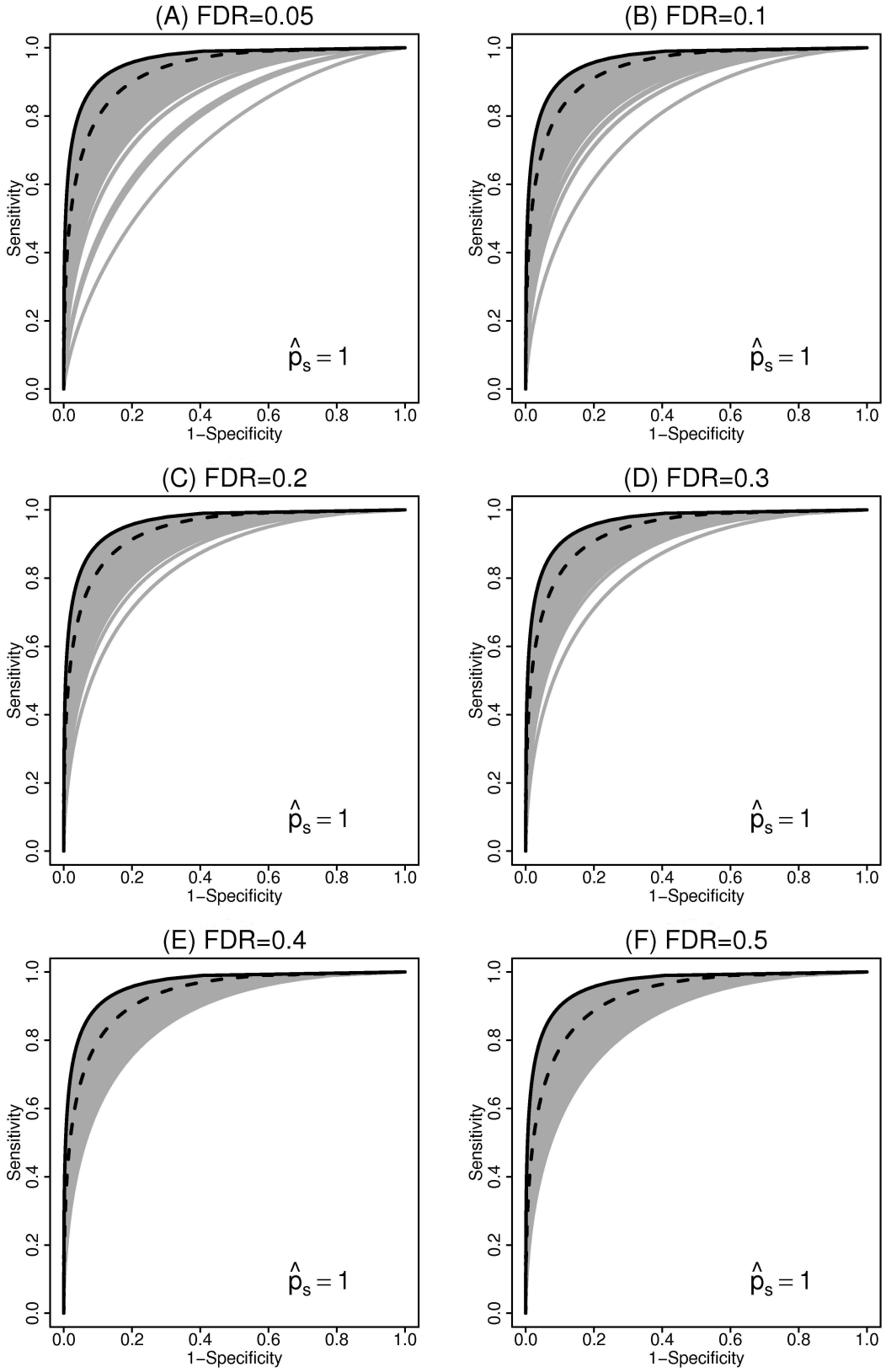

Figure 4.3: Simulation results for the protected approach: simulated ROC-curves (1000 plotted curves out of 10000 simulation steps) for a future patient for different FDR-values. 10 among 1000 hypotheses are assumed to be alternatives. The sample size per group was set to $n = 50$. The average curve (dashed line) and the theoretically best ROC-curve (solid line) is given.

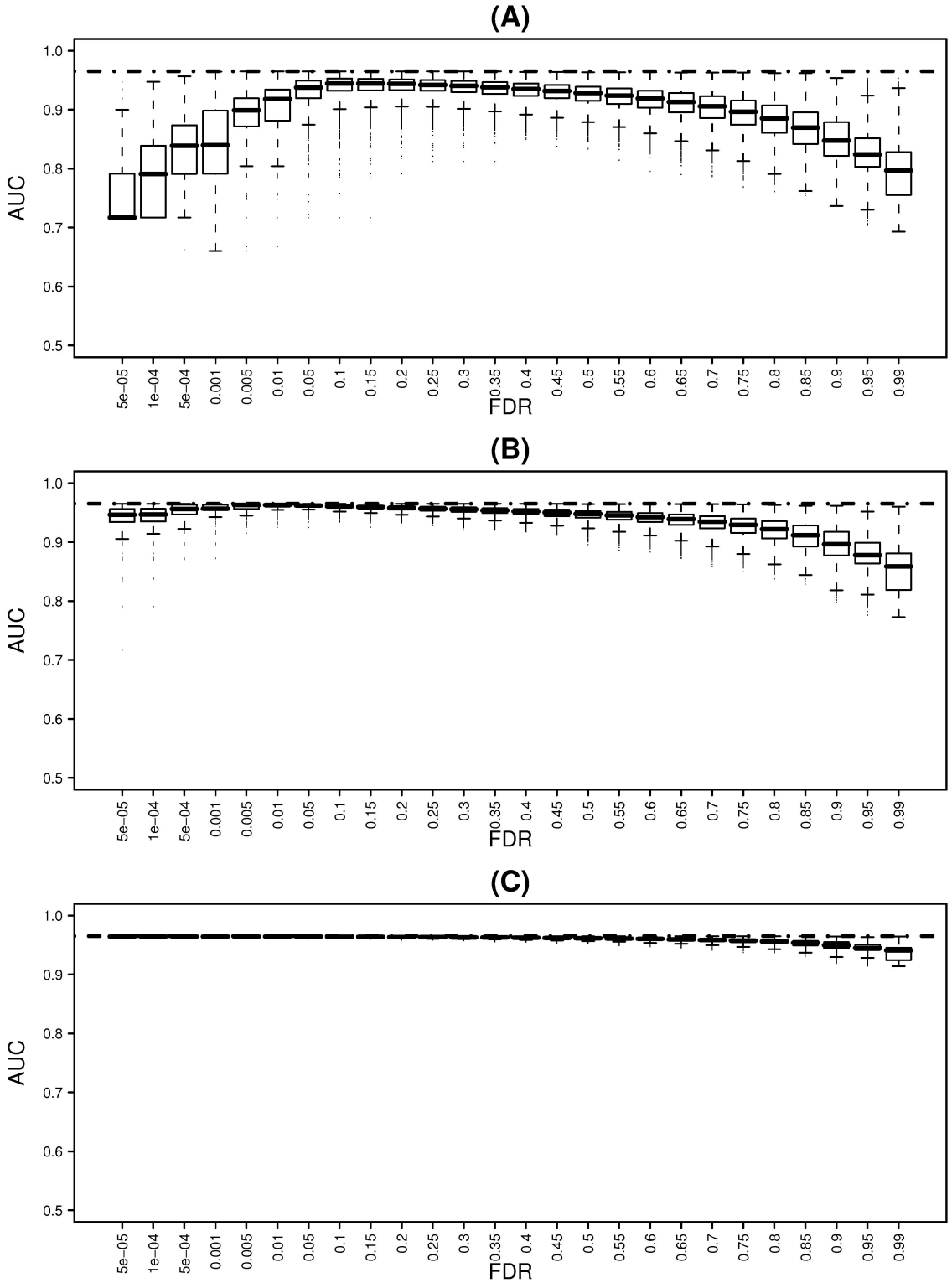

Figure 4.4: Boxplots of the area under the ROC-Curve for selection using the protected approach assuming $m_e = 10$ alternatives markers among $m = 1000$ tested markers (10000 simulation steps). The sample size per group was set to $n = 50$ (Figure (A)), 100 (Figure (B)) and 500 (Figure (C)). AUC_* is given as dotdashed horizontal line. Δ was set to 0.81.

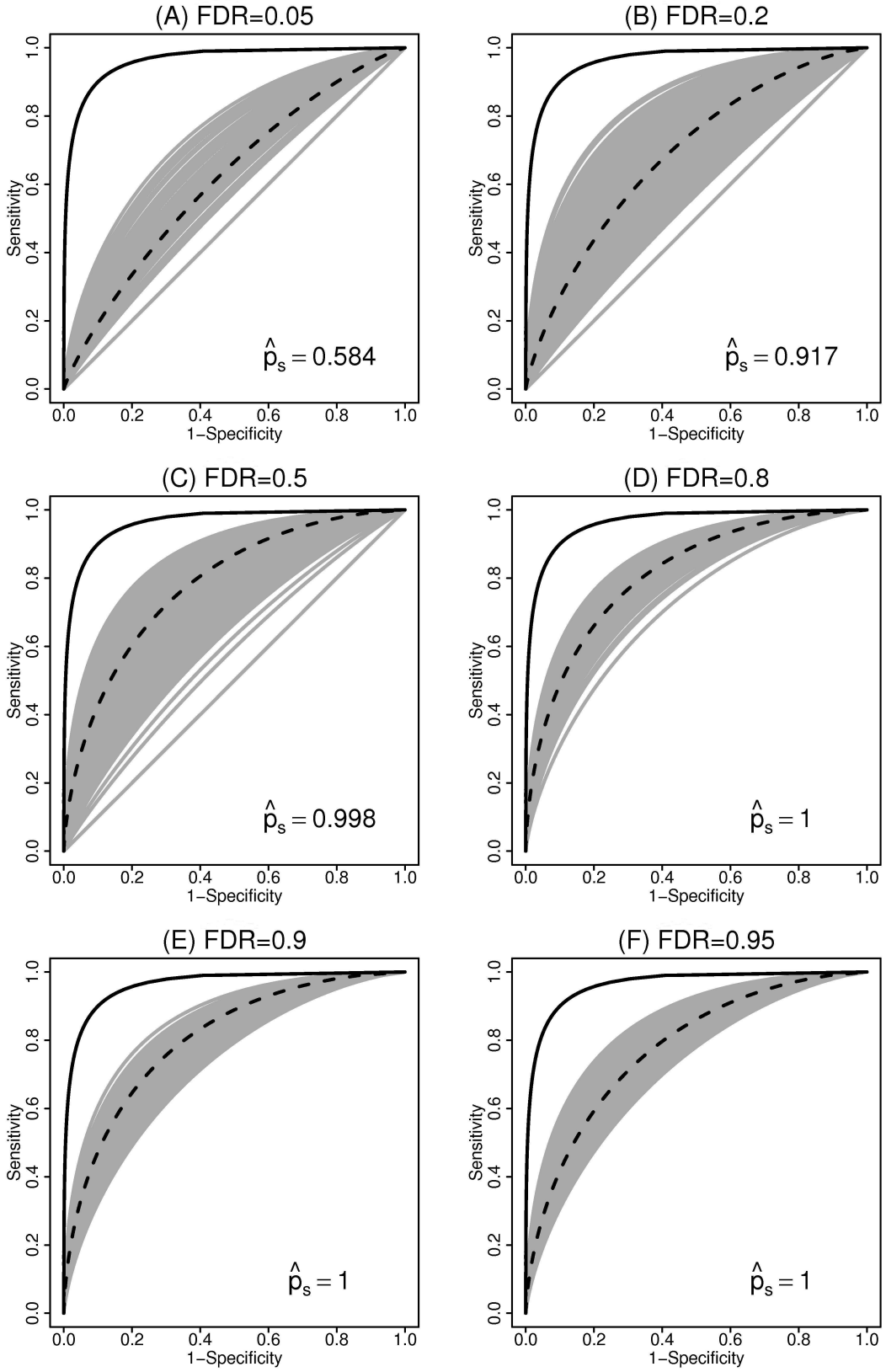

Figure 4.5: Simulation results for the protected approach: simulated ROC-curves (1000 plotted curves out of 10000 simulation steps) for a future patient for different FDR-values. 60 among 1000 hypotheses are assumed to be alternatives. The sample size per group was set to $n = 50$. The average curve (dashed line) and the theoretically best ROC-curve (solid line) is given.

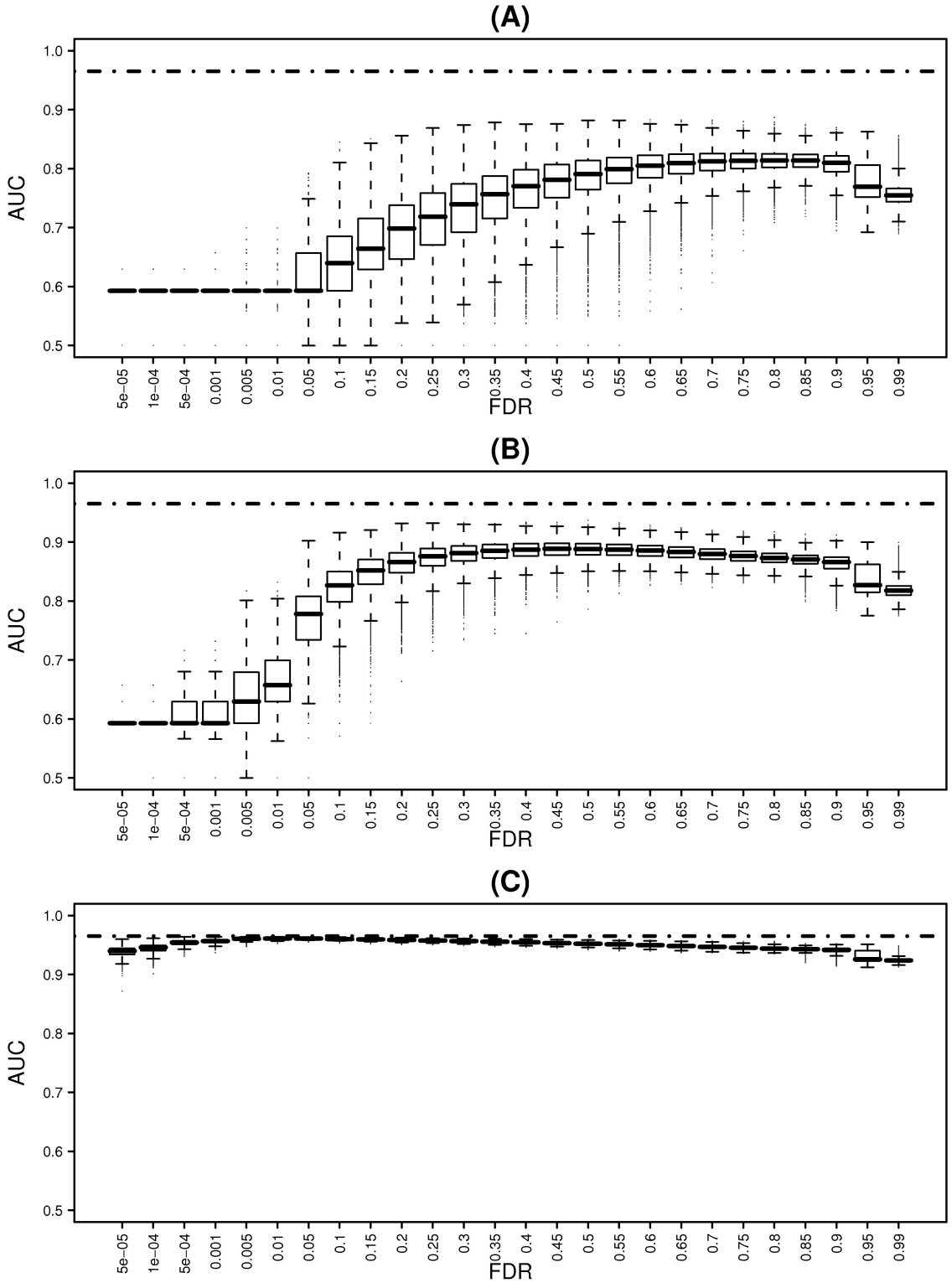

Figure 4.6: Boxplots of the area under the ROC-Curve for selection using the protected approach assuming $m_e = 60$ alternatives markers among $m = 1000$ tested markers (10000 simulation steps). The sample size per group was set to $n = 50$ (Figure (A)), 100 (Figure (B)) and 500 (Figure (C)). AUC_* is given as dotdashed horizontal line. Δ was set to 0.33.

m_e=10, m=6000

Let us furthermore have a look at the situation assuming $m_e = 10$ effective markers (alternatives) among $m = 6000$ hypotheses. Because of the larger number of tested hypotheses the problem to find the alternatives becomes harder. The effect size Δ remains the same, thus $\Delta = 0.81$ for $m_e = 10$ since it only depends on the number m_e of alternatives and not on the number of tested hypotheses. Figure 4.7 (A) shows the situation fixing the sample size per group to $n = 50$. Clearly the average AUC values for the grid of investigated FDRs are smaller as compared to the scenario with m only equal to 1000. Due to the large effect size assuming $m_e = 10$, selection controlling a FDR of 0.25 leads to good prediction scores with an AUC_*^{sim} of 0.918, although selecting out of 6000 hypotheses. Increasing the sample size to $n = 100$ per group (Figure 4.7 (B)) $AUC_*^{sim} = 0.959$ occurs for FDR= 0.05. The results of a further increase of the sample size per group to $n = 500$ can be seen in Figure 4.7 (C). Again, as in the case of $m = 1000$, because of the good estimates of the group means and the very small p-values of the alternatives, for a wide range of FDR values the future performance remains good. Up to a FDR of 0.01, AUC_*^{sim} for future prediction is almost equal to the best theoretically achievable AUC_*. For larger FDR values the performance is only slightly smaller. $\hat{p}_s = 1$ in the simulations for all FDR selection criteria if $n = 500$ per group.

m_e=60, m=6000

Assuming $m_e = 60$ among $m = 6000$ hypotheses, $\Delta = 0.33$. Now the situation gets worse. Because of the small effect size and the larger number of hypotheses to test no good prediction score can be selected if the sample size per group is small. Figure 4.8 (A) shows the results of the simulations when fixing the sample size per group to $n = 50$. When choosing FDR= 0.9 as selection criterion on average only 28.7 out of the 60 alternatives and additionally a mean value of 350.5 true null hypotheses are selected which leads to prediction scores achieving $AUC_*^{sim} = 0.669$. Hence, AUC_*^{sim} is much smaller than $AUC_* = 0.965$. This again describes exactly the problem of gene expression studies. If only a few effective

markers with a large effect size exist it may be possible to find good prediction scores if the right selection criterium is used. In contrast, if there are many markers with low effect sizes working together, searching for prediction scores with rather small sample sizes becomes a formidable problem. Note also that if selection is based on a FDR of 0.05 in only 26% of the simulated samples a prediction score is selected from the data. Even if the FDR is 0.85 it happens that no prediction score is constructed from the data ($\hat{p}_s = 0.995$).

Increasing the sample size to $n = 100$ per group (Figure 4.8 (B)) the situation improves a little as compared to $n = 50$. However the best performance occurs for FDR= 0.6 with $AUC_*^{sim} = 0.792$. Then 26.7 out of the 60 alternatives and 43.1 true null hypotheses are selected for future prediction. In the simulation studies $\hat{p} = 1$ if the FDR is larger than 0.40. A further increase of the sample size per group to $n = 500$ changes the situation completely (see Figures 4.8 (C)). Again, for a wide range of FDR values the future performance remains good. The largest average AUC among the investigated FDR values (AUC_*^{sim}) occurs for FDR= 0.05. In average 58.5 effective and 3.1 ineffective markers are selected which lead to prognostic scores with $AUC_*^{sim} = 0.959$. $\hat{p}_s = 1$ in the simulations for all FDR selection criteria if $n = 500$ per group.

For more detailed information see Tables E.8 to E.13 in the appendix.

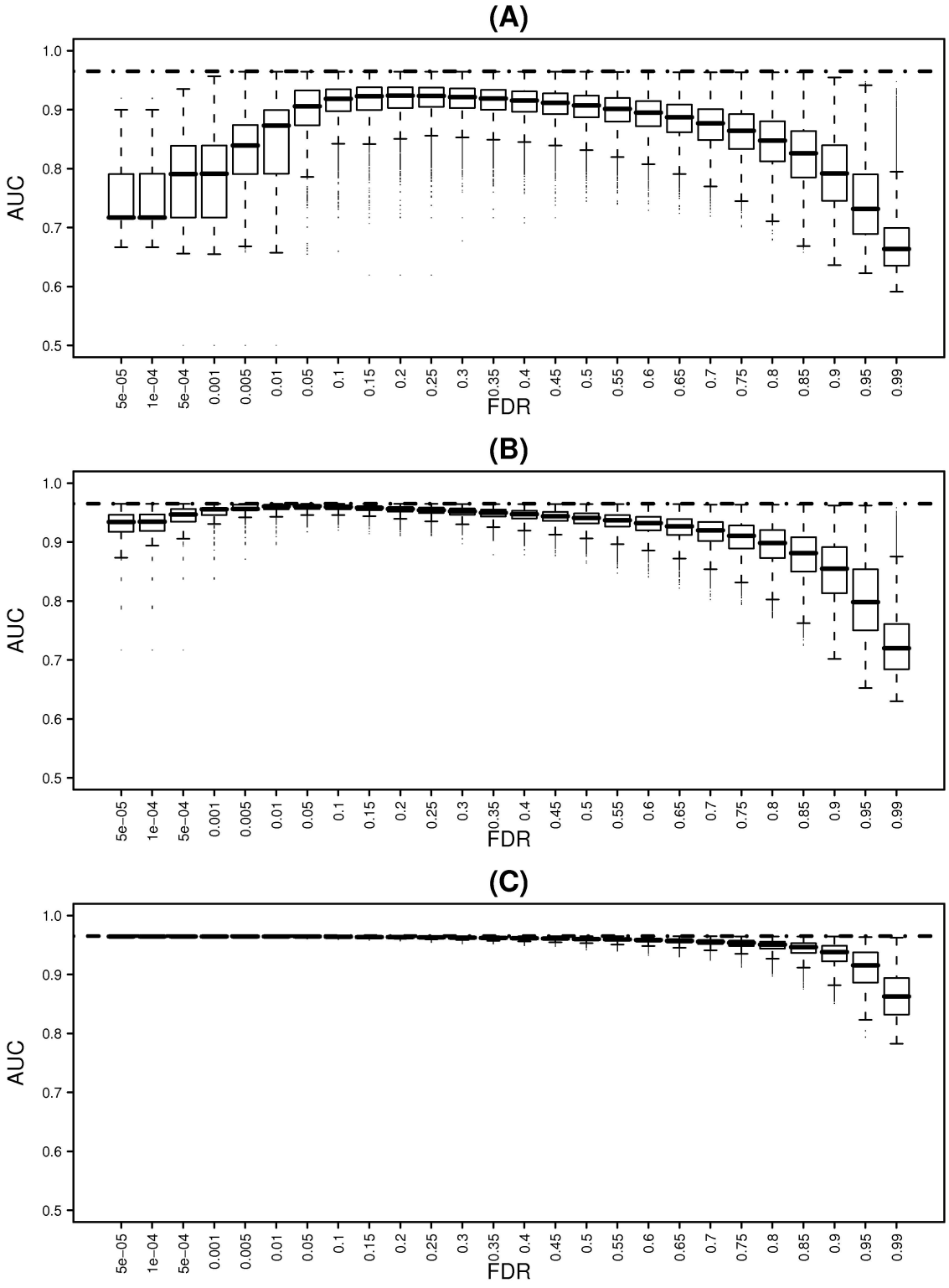

Figure 4.7: Boxplots of the area under the ROC-Curve for selection using the protected approach assuming $m_e = 10$ alternatives markers among $m = 6000$ tested markers (10000 simulation steps). The sample size per group was set to $n = 50$ (Figure (A)), 100 (Figure (B)) and 500 (Figure (C)). AUC_* is given as dotdashed horizontal line. Δ was set to 0.81.

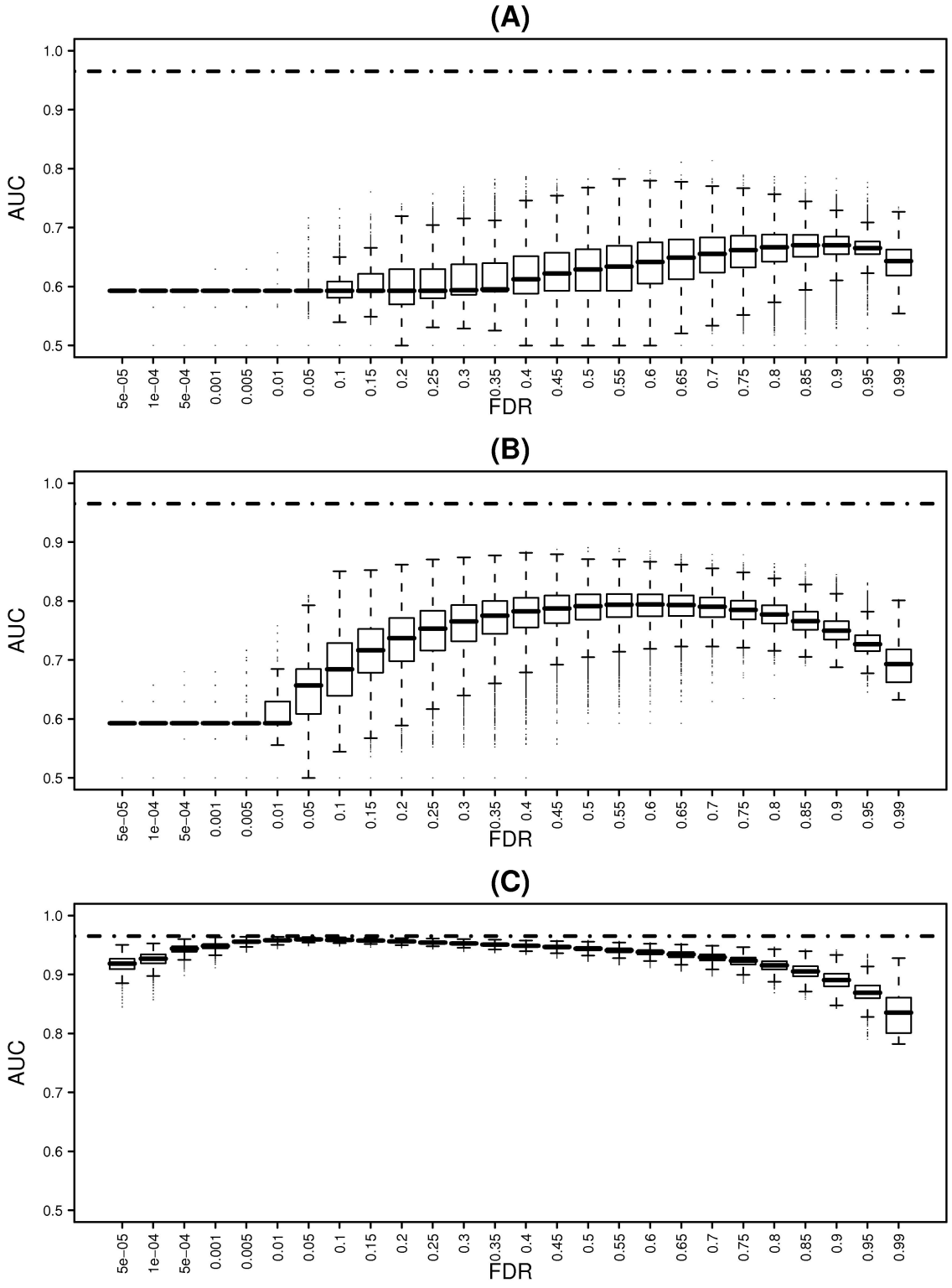

Figure 4.8: Boxplots of the area under the ROC-Curve for selection using the protected approach assuming $m_e = 60$ alternatives markers among $m = 6000$ tested markers (10000 simulation steps). The sample size per group was set to $n = 50$ (Figure (A)), 100 (Figure (B)) and 500 (Figure (C)). AUC_* is given as dotdashed horizontal line. Δ was set to 0.33.

4.6.2 Variable selection using forward logistic regression

For selection using the forward logistic regression, again Δ was set to 0.81 in the case of $m_e = 10$ and to 0.33 for $m_e = 60$. Because of run-time problems, simulations (1000 simulation steps) were only performed for $m = 1000$. Different thresholds were applied for the individual p-values in the stepwise selection based on the multiple logistic regression. Those thresholds γ where determined corresponding to the rejection boundary when controlling the FDR of 0.02, 0.05, 0.2 and 0.3 in a conventional single-stage design. Thus for $m_e = 10$ we used $\gamma = 0.0002$, 0.0005, 0.0025 and 0.0043 and for $m_e = 60$, $\gamma = 0.0001$, 0.0006, 0.0073 and 0.0015. We additionally performed simulations using the Bonferroni correction to calculate the threshold ($\gamma = 0.00005$) as well as the situation were no correction ($\gamma = 0.05$) is performed.

Figures 4.9 and 4.10 show the ROC-curves (grey curves), the average ROC (dashed line) and the theoretically best ROC (solid line) for the different γ values for the situation of $m_e = 10$ and 60 respectively. The simulation results of the forward logistic regression show that this selection method for the investigated scenarios performs poor in terms of AUC values for prediction of the outcome of future patients as compared to the selection procedure using the FDR approach. Figure 4.11 shows boxplots of the AUC values for future prediction using the forward logistic regression for the different γ values for $m_e = 10$ (first column) and $m_e = 60$ (second column) among $m = 1000$ hypotheses considering a sample size of $n = 50$ (first row) and $n = 100$ per group (second row). The poor result may be due to the following reason: in a small training sample the forward logistic regression leads to a complete separation of data points, i.e. responders and non-responders of the validation data set can be fully separated with the found regression model. Only a few effective markers are selected for future prediction using the forward logistic regression.

The best performance for the situation of $m_e = 10$ occurs for $\gamma = 0.0005$ with $AUC_*^{sim} = 0.812$ for $n = 50$ and again at $\gamma = 0.0005$ with $AUC_*^{sim} = 0.900$ for $n = 100$. The forward logistic regression applying a larger sample size clearly performs better. However, AUC_*^{sim} for

the selection procedure using the FDR was 0.941 for $n = 50$ and 0.961 for $n = 100$. Note that for $n = 50$ in average only 2.881 and for $n = 100$ only 5.644 out of the 10 alternatives are selected using the logistic regression.

In the case of $m_e = 60$ for $\gamma = 0.0073$ up to 0.05 the mean AUC on average is 0.613 for $n = 50$. Because of complete separation of data points for larger γ values, the same performance is achieved for $\gamma \geq 0.0073$. Setting $n = 100$ per group a similar result can be seen. Again choosing $\gamma \geq 0.0073$ leads to the same performance achieving $AUC_*^{sim} = 0.696$. Note that AUC_*^{sim} for selection using the FDR was 0.813 for $n = 50$ and 0.887 for $n = 100$ per group. However, the theoretically achievable AUC_* is 0.965, which is not achieved with both selection methods. Note also that if γ was chosen to be 0.00005, only in 26% of the simulated samples a prediction score is selected from the data if $n = 50$ and in 86.3% if $n = 100$ per group.

For more detailed information see Table E.7 to in the appendix.

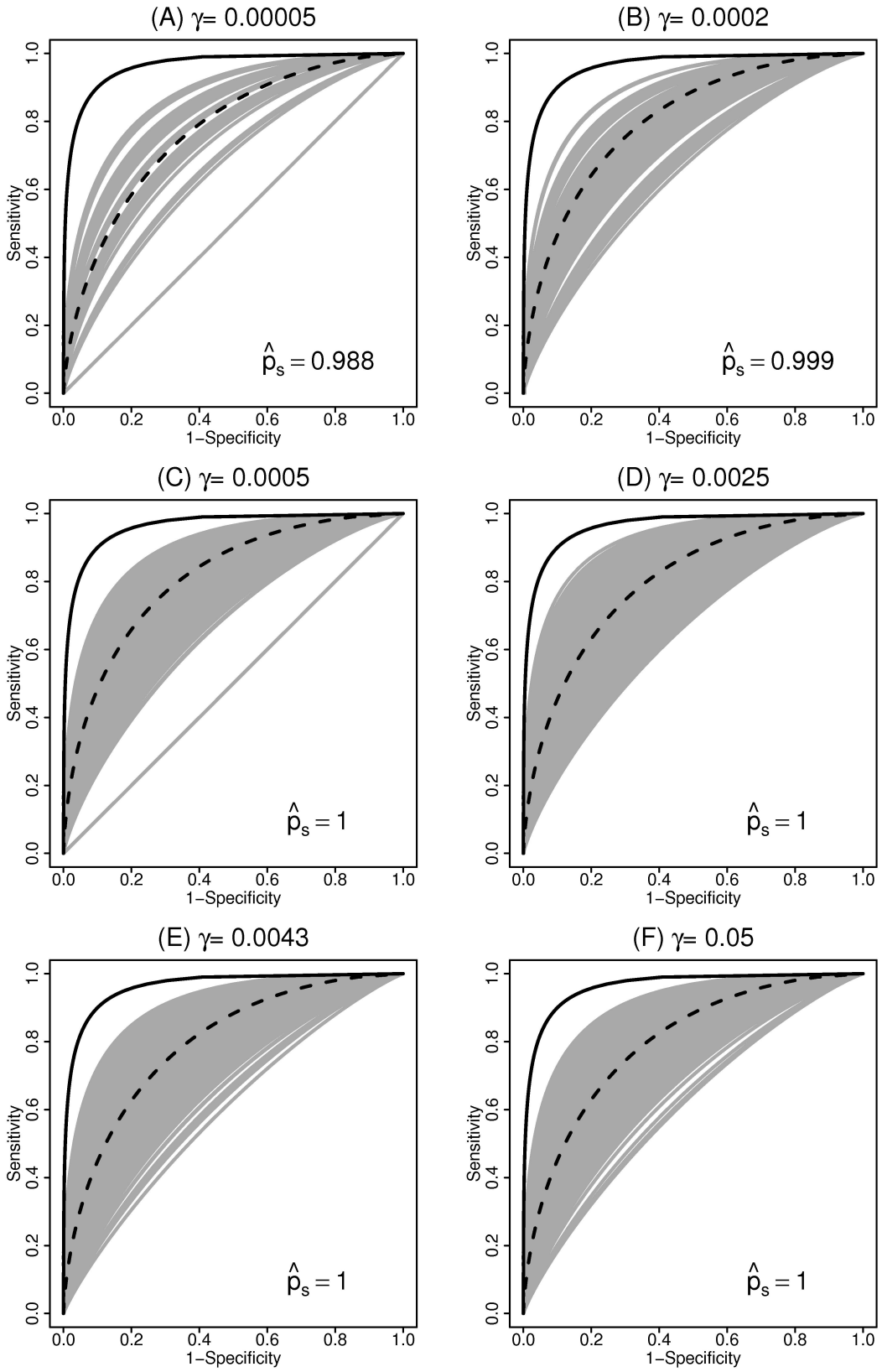

Figure 4.9: Simulation results using forward logistic regression (1000 steps): simulated ROC-curves for a future patient for different values of γ. 10 among 1000 hypotheses are assumed to be alternatives. The sample size per group was set to $n = 50$. The average curve (dashed line) and the theoretically best ROC-curve (solid line) is given.

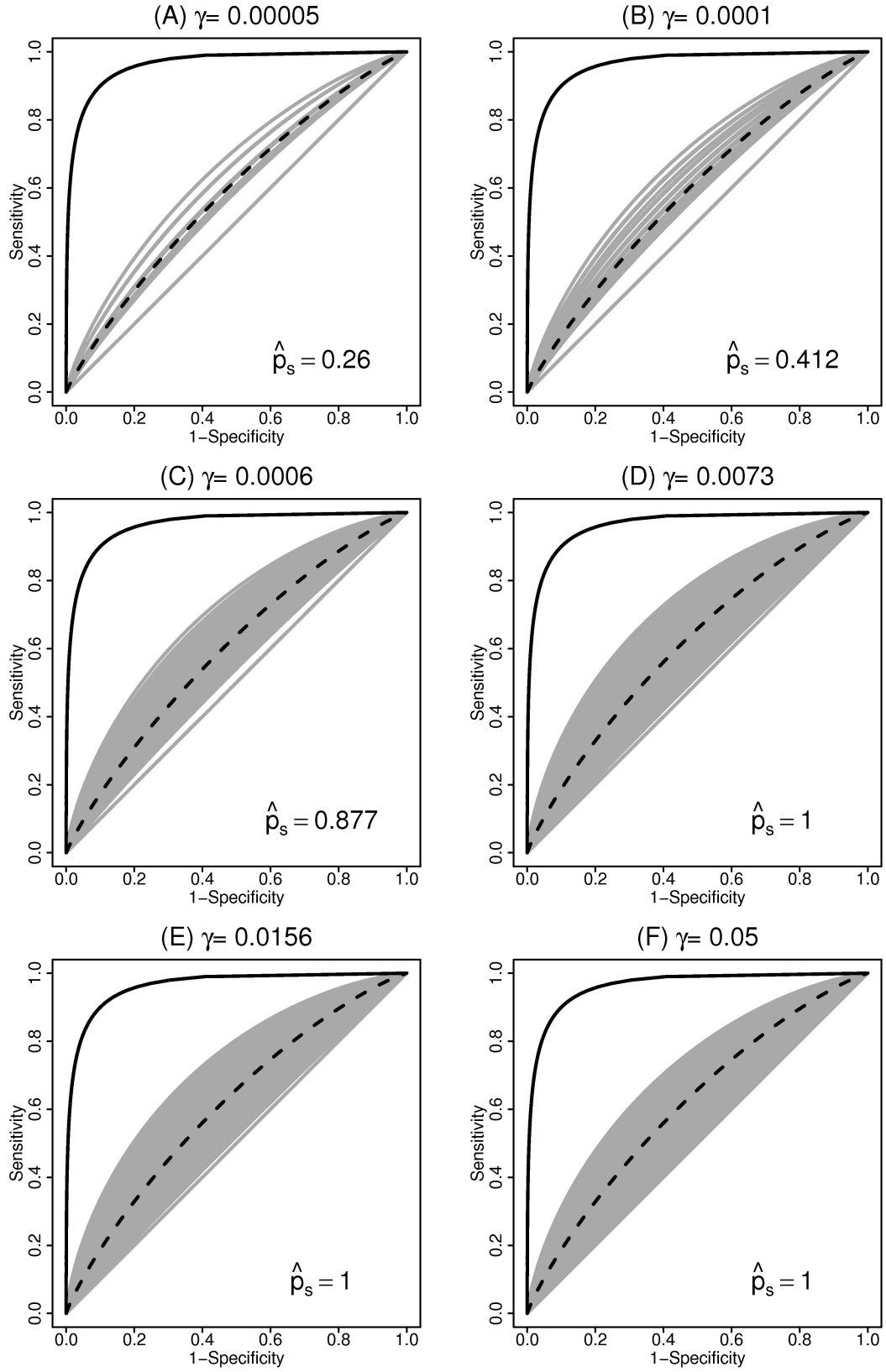

Figure 4.10: Simulation results using forward logistic regression (1000 steps): simulated ROC-curves for a future patient for different values of γ. 60 among 1000 hypotheses are assumed to be alternatives. The sample size per group was set to $n = 50$. The average curve (dashed line) and the theoretically best ROC-curve (solid line) is given.

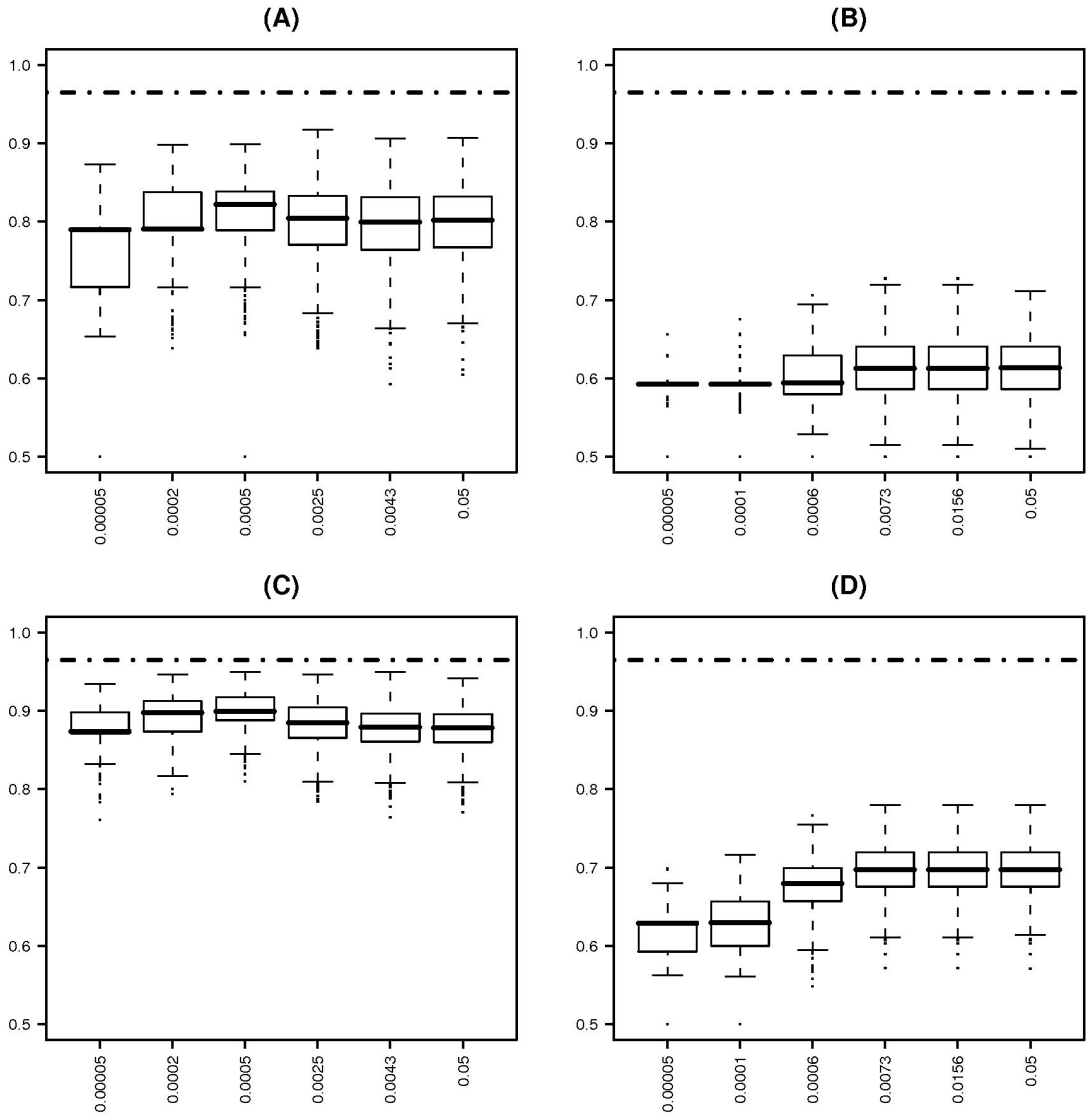

Figure 4.11: Boxplots of the area under the ROC-Curve for 1000 simulated samples for selection using forward logistic regression assuming $m_e = 10$ (first column) and $m_e = 60$ (second column) effective markers among $m = 1000$ tested markers and assuming a sample size of $n = 50$ (first row) and $n = 100$ (second row). AUC_* is shown as dotdashed horizontal line. Results for different γ values are given.

4.6.3 Variable selection using the optimistic approach

The effect size is again set to $\Delta = 0.81$ if $m_e = 10$ and $\Delta = 0.33$ if $m_e = 60$, thus forcing the optimal ROC-curve through the point where sensitivity and specificity are equal to 0.9. Note that for the optimistic approach selecting the best k markers, $p_s \equiv 1$, thus always a prediction score is selected from the data.

m=1000

Figures 4.12 show simulated ROC-curves (grey curves) for some k assuming $m_e = 10$ among $m = 1000$ hypotheses. Again the average ROC calculated from the samples (dashed curve) and the theoretically best curve (solid curve) are given. Figure 4.13 (A) shows boxplots of the future AUC values for different values of k fixing the sample size per group $n = 50$. $AUC_*^{sim} = 0.942$ occurs when selecting the best 10 hypotheses. In average 8.6 effective and thus 1.4 ineffective markers are included in the prediction score. When increasing the sample size per group to $n = 100$ (Figure 4.13 (B)) again the best performance among the scenarios considered in the simulations is achieved for $k = 10$ with an AUC_*^{sim} of 0.962. As for the FDR-selection criterion for larger sample sizes the mean values for the group of responders and non-responders can be better estimated and thus also for larger values of k the performance still remains good. This can be also seen in the case of $n = 500$ (see Figure 4.13 (C)).

Figures 4.14 show simulated ROC-curves (grey curves) as well as the average curve (dashed line) and the theoretically best ROC-curve (solid line) for some k assuming $m_e = 60$ among $m = 1000$ hypotheses. Figure 4.15 (A) shows boxplots of the future AUC values for different values of k for this scenarios for a sample size per group of $n = 50$. Now the optimum is very flat. A similar performance with an AUC_*^{sim} between 0.813 and 0.814 can be achieved varying k from 150 to 600. If k is set to 600, in average 541.9 true null hypotheses and 58 alternatives are selected. However, the large number of markers may result in a unfeasible score. If selecting the best 150 hypotheses, in average 40.6 alternatives and 109.4 true null hypotheses are selected which lead to a similar average future AUC. Choosing $k = m_e = 60$

on average 33.5 true null hypotheses but only 26.5 out of the 60 alternatives are selected for a future prediction which results in an average AUC of only 0.80. Fixing the sample size to $n = 100$ (Figure 4.15 (B)), among the considered scenarios, $AUC_*^{sim} = 0.888$ is achieved for $k = 100$. Again only a slight decrease of the mean AUC can be seen when using a larger k (see also Figure 4.15 (C) for a sample size of $n = 500$). Note that a large decrease of the mean AUC occurs if k is smaller than m_e, as in the case of $k = 5$ where the mean future AUC is only 0.695 assuming a sample size of $n = 100$ per group.

m=6000

Figures 4.16 (A) and 4.17 (A) show the situation when increasing the number of tested hypotheses to 6000. The result is similar as for selection using the FDR approach. If the number of alternatives is small, but their effect size is large one may get a good prediction score even if the sample size is small. For larger sample sizes, the estimates of the group means and thus of the weights in the score get more precisely. For a larger range of k the performance remains good. If a larger number of alternatives is expected with a rather small effect size one may not get a good prediction score applying a small sample size. Assuming $m_e = 60$ and applying a samples size of $n = 50$ per group, the average AUC values over all investigated k are not exceeding 0.7. For a larger sample size per group the situation clearly gets better, for a sample size of $n = 500$ per group, the best performance occurs if $k = 60 (= m_e)$ achieving an AUC_*^{sim} of 0.959.

For more detailed information see also Tables E.1 to E.6 in the appendix.

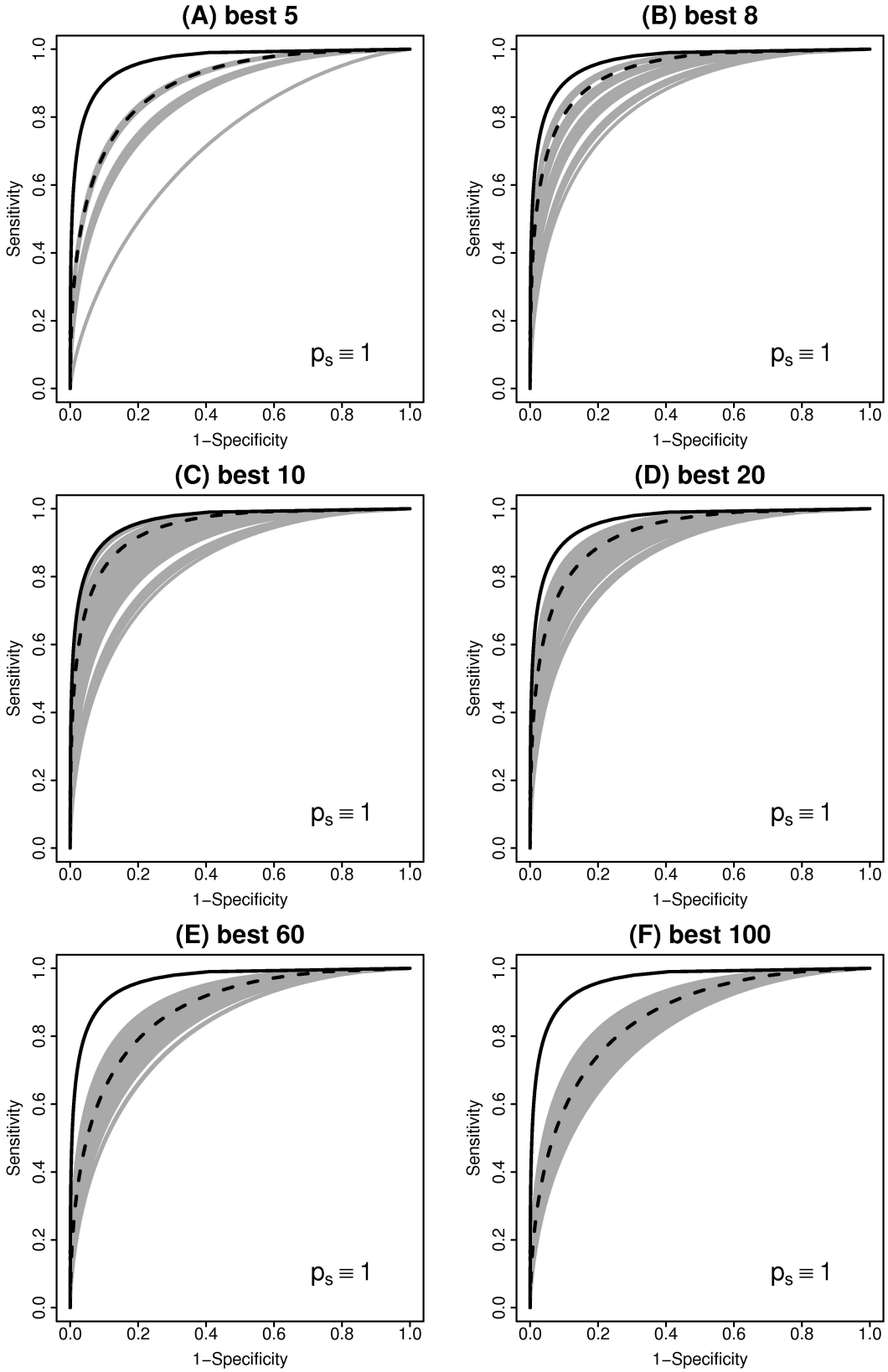

Figure 4.12: Simulation Results for the protected approach: simulated ROC-curves (1000 plotted curves out of 10000 simulation steps) for a future patient for different values of k. 10 among 1000 hypotheses are assumed to be alternatives. The average curve (dashed line) and the theoretically best ROC-curve (solid line) is given.

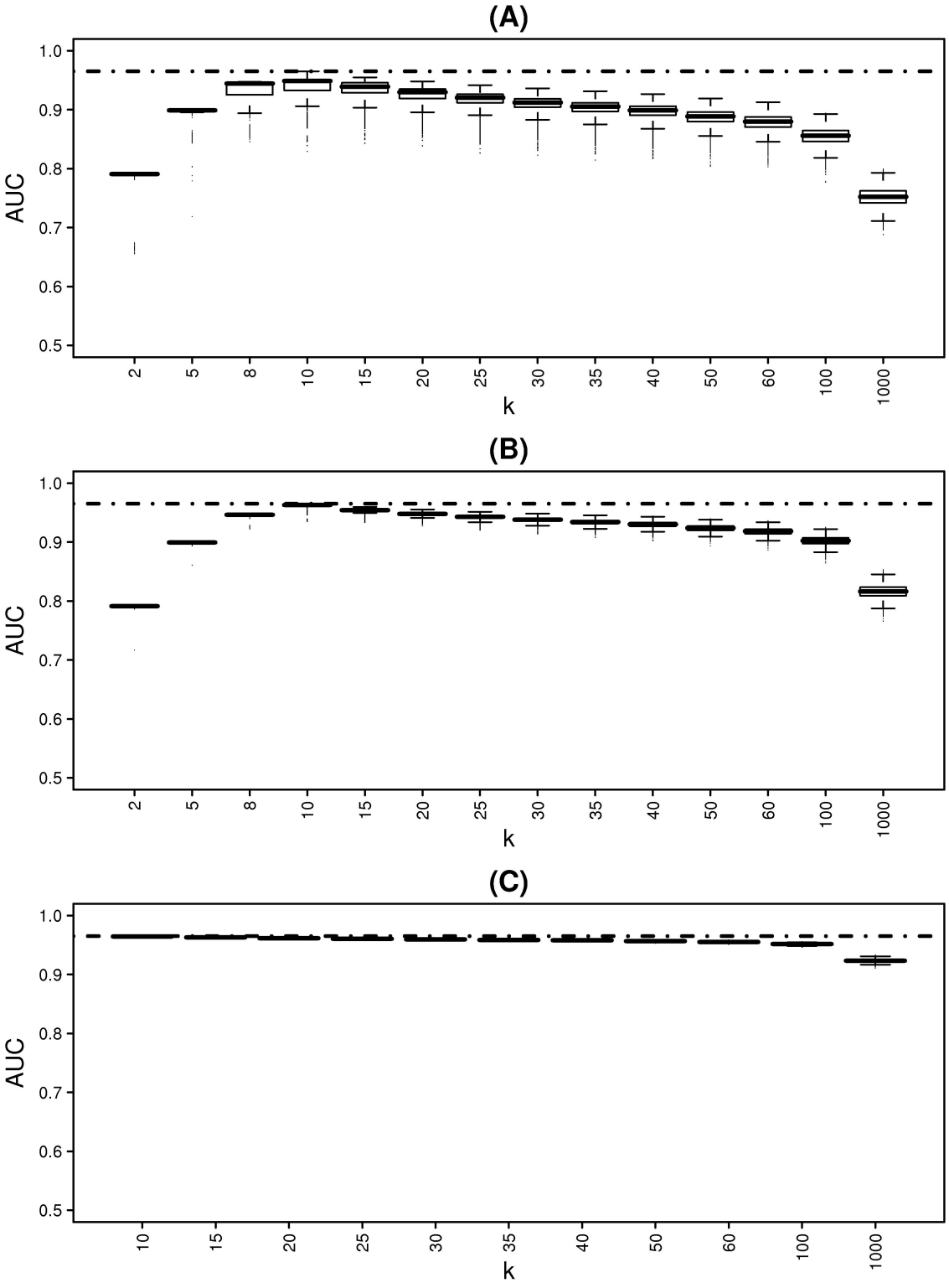

Figure 4.13: Boxplots of the area under the ROC-Curve for selection using optimistic approach assuming $m_e = 10$ alternatives markers among $m = 1000$ tested markers (10000 simulated samples). The sample size per group was set to $n = 50$ (Figure (A)), 100 (Figure (B)) and 500 (Figure (C)). AUC_* is given as dotdashed horizontal line. Δ was set to 0.81.

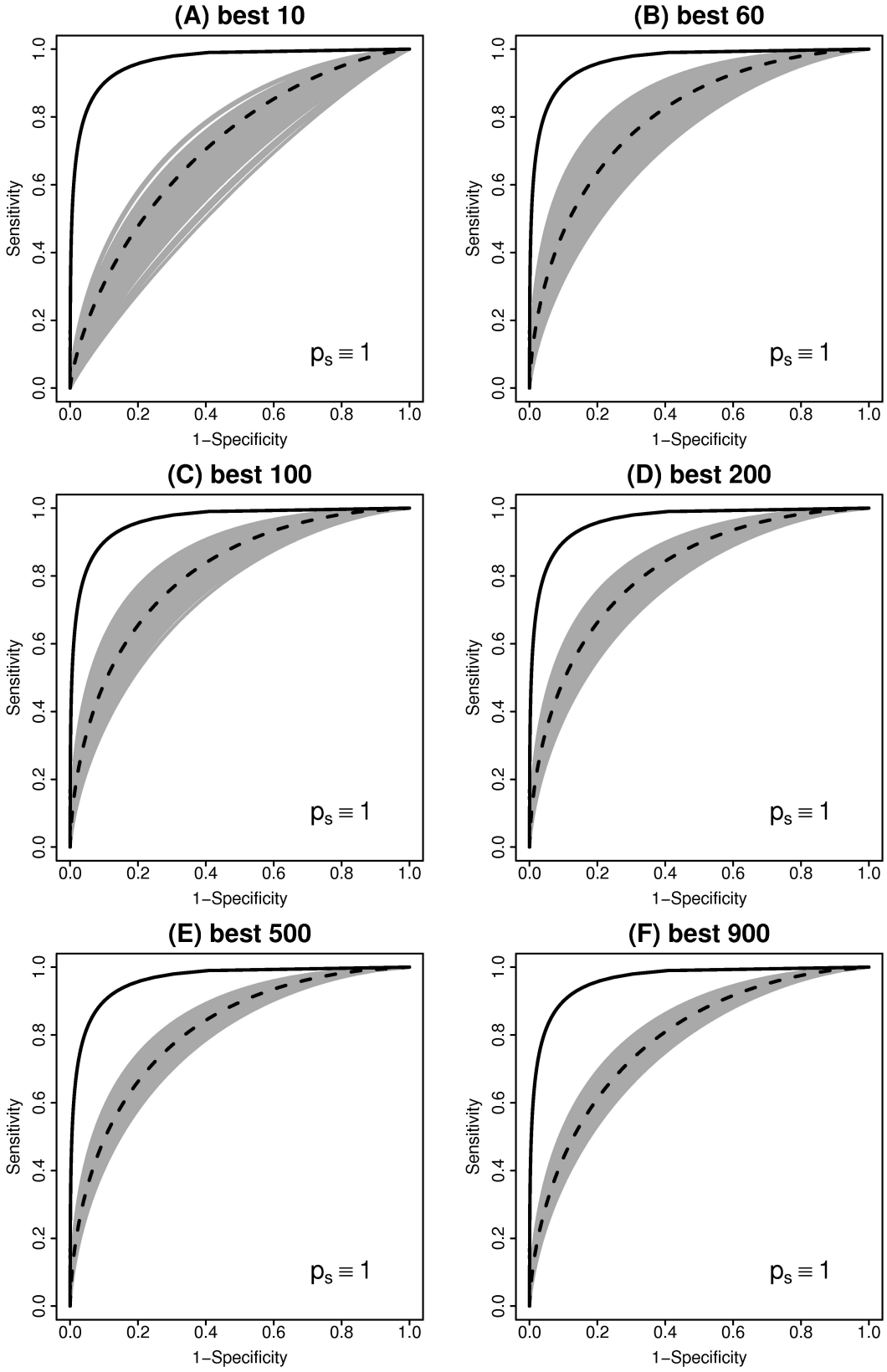

Figure 4.14: Simulation Results for the protected approach: simulated ROC-curves (1000 plotted curves out of 10000 simulation steps) for a future patient for different values of k. 60 among 1000 hypotheses are assumed to be alternatives. The average curve (dashed line) and the The theoretically best ROC-curve (solid line) is given.

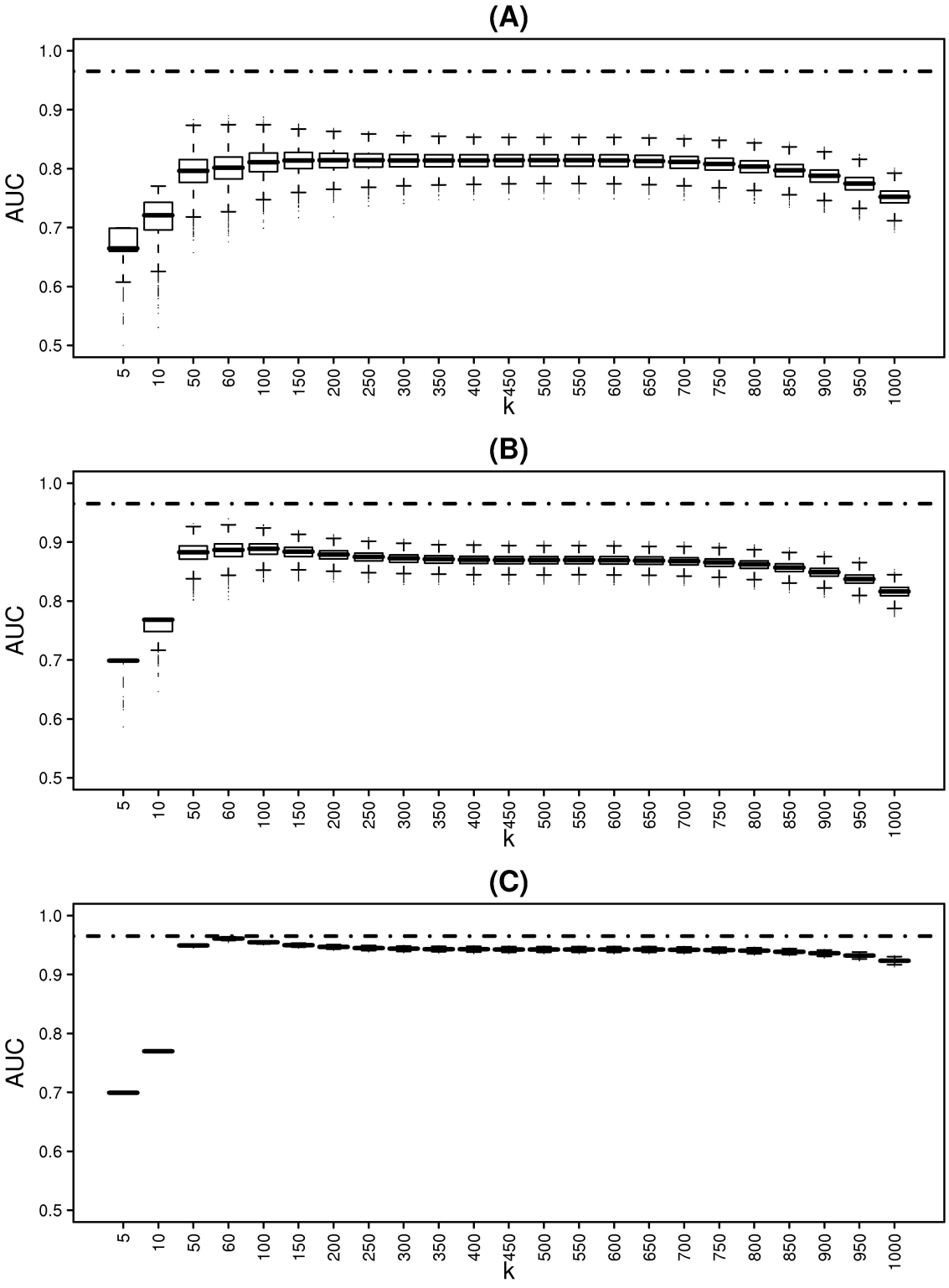

Figure 4.15: Boxplots of the area under the ROC-Curve for selection using optimistic approach assuming $m_e = 60$ alternatives markers among $m = 1000$ tested markers (10000 simulated samples). The sample size per group was set to $n = 50$ (Figure (A)), 100 (Figure (B)) and 500 (Figure (C)). AUC_* is given as dotdashed horizontal line. Δ was set to 0.33.

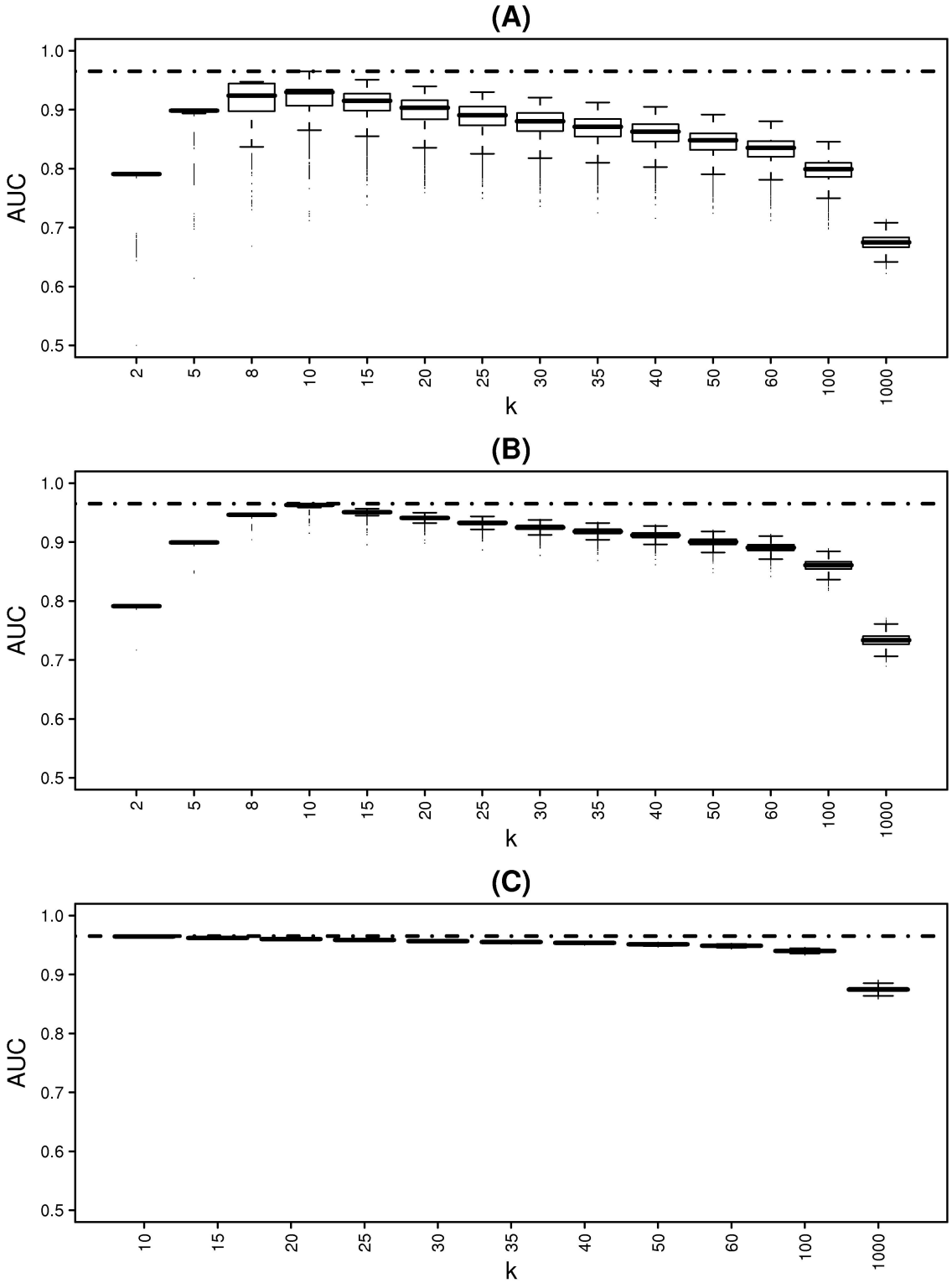

Figure 4.16: Boxplots of the area under the ROC-Curve for selection using optimistic approach assuming $m_e = 10$ alternatives markers among $m = 6000$ tested markers (10000 simulated samples). The sample size per group was set to $n = 50$ (Figure (A)), 100 (Figure (B)) and 500 (Figure (C)). AUC_* is given as dotdashed horizontal line. Δ was set to 0.81.

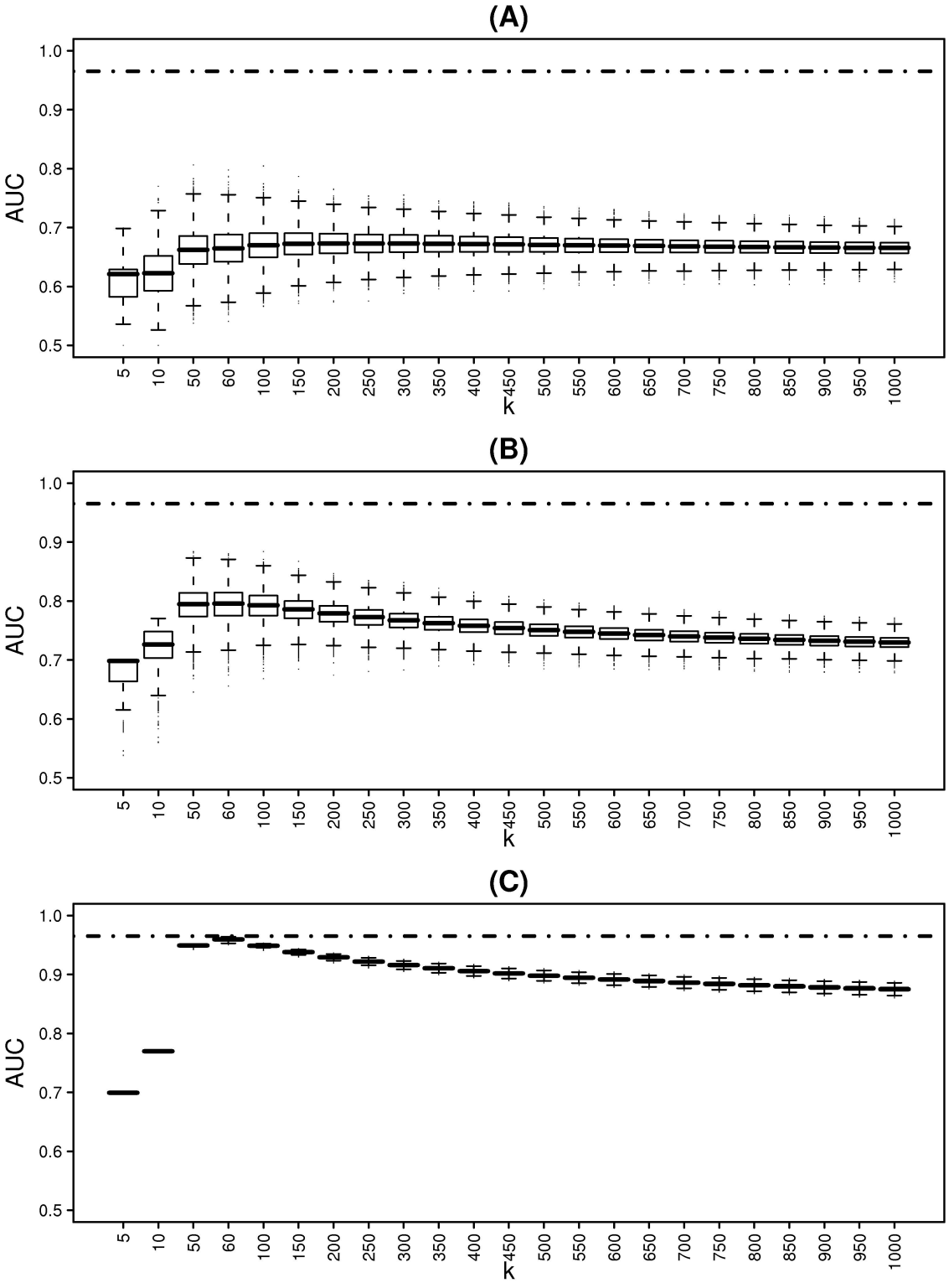

Figure 4.17: Boxplots of the area under the ROC-Curve for selection using optimistic approach assuming $m_e = 60$ alternatives markers among $m = 6000$ tested markers (10000 simulated samples). The sample size per group was set to $n = 50$ (Figure (A)), 100 (Figure (B)) and 500 (Figure (C)). AUC_* is given as dotdashed horizontal line. Δ was set to 0.33.

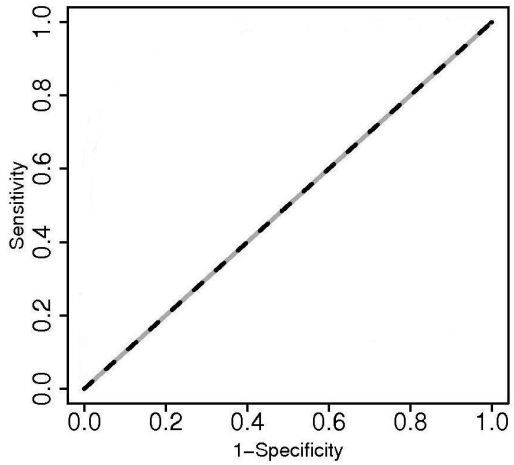

Figure 4.18: ROC-Curves: Situation under global null hypothesis

4.6.4 Situation under the global null hypothesis

Figure 4.18 shows the situation under the global null hypothesis of no existing effective markers at all ($m_e = 0$). The ROC curve is always the diagonal (AUC= 0.5). Whatever selection procedure is used, if markers are selected, they are always true null hypotheses and thus the prediction score is useless. For the protected procedure, i.e. selection using the FDR approach, by definition the probability to end with a selection of markers and building a score is targeted at the pre-chosen FDR. Hence in the case of the global null hypotheses it would have been better to chose a small FDR. However, if a large number of alternatives are expected with rather small effect sizes a very large FDR should be chosen as selection criterion. The unprotected procedure, selecting the k best markers for prediction, leads to a completely non-informative score in all cases (per definition, $p_s \equiv 1$).

If we select markers using the forward logistic regression, as for the FDR selection method, the results depend on the boundary γ chosen a priori. E.g. if we decided for Bonferroni corrected boundaries in only 2.9% of the 1000 simulated cases at least one marker was identified for future prediction but when increasing γ to 0.0025 in 98.3% useless prediction scores are produced. Without correction ($\gamma = 0.05$) $\hat{p}_s = 1$.

This again demonstrates the dilemma of the task we are faced with. It may be possible

to improve the selection and estimation procedures but a contradiction will remain: being cautious may help not to produce too many nuisance results if the postulated relationships do not exist. Being more optimistic and liberal may improve the results, particularly if the true state of a nature is close to what the selection procedure is targeted at, e.g., if in our case k is close to the actual number of effective markers. However the unprotected procedure is vulnerable for the situations of no existing effects. Here it will always produce nuisance results and due to a large number of tests the magnitude of the observed best effects may mislead the experimenter.

4.7 Estimating the selection criterion using jackknife

4.7.1 Jackknife for the protected approach

We have seen from the previous sections that the forward logistic regression may not be a good selection procedure if small sample sizes are given. There are simple methods, like using the FDR approach and just building a weighted sum of the selected hypotheses, which, if the right selection boundary is chosen, lead to a better performance in terms of AUC values for prediction of the outcome of a future patient. However, it remains the question, how to choose the selection boundary, because depending on the parameter constellation (varying number of tested hypotheses, varying proportion of effective markers and varying sample size), different boundaries are required in order to achieve a large AUC for future independent patients.

The procedure

To estimate an appropriate selection boundary γ for the protected method we will now investigate a modified Jackknife procedure. The marker levels are again assumed to follow independent normal distributions with known variance $\sigma^2 = 1$.

From a given data set with n responders and n non-responders, a pair of a single responder and a single non-responder respectively is left out. Note that there are $j = 1, ..., n^2$ possibilities for leaving out a pair of one responder and one non-responder. The $2(n-1)$ patients in each of the n^2 "training" samples respectively are used to estimate prediction scores applying a grid of values γ_i, $i = 1, ...s$, for the selection boundary. As discussed in Section 4.3.1, the markers, whose one sided p-values lie below the selection boundary γ_i are selected for future prediction. For the left out responder and non-responder respectively we now calculate for each γ_i, $i = 1, ...s$, the value of the corresponding prediction score:

$$\hat{f}_j(\mathbf{x}_{r,j}; \gamma_i) = (\hat{\boldsymbol{\mu}}_{r,(j)}(\gamma_i) - \hat{\boldsymbol{\mu}}_{nr,(j)}(\gamma_i))^T \mathbf{x}_{r,j}$$

and

$$\hat{f}_j(\mathbf{x}_{nr,j};\gamma_i) = (\hat{\boldsymbol{\mu}}_{r,(j)}(\gamma_i) - \hat{\boldsymbol{\mu}}_{nr,(j)}(\gamma_i))^T \mathbf{x}_{nr,j}$$

where $\hat{\boldsymbol{\mu}}_{r,(j)}(\gamma_i)$ and $\hat{\boldsymbol{\mu}}_{nr,(j)}(\gamma_i)$ are the mean values of the group of responders and non-responders respectively of the markers selected for prediction, calculated from the jth training sample, using γ_i as selection boundary. The notation $\hat{\boldsymbol{\mu}}_{a,(j)}(\gamma_i)$, $a = r, nr$, indicates that the length of the vector is nondecreasing in γ_i; $\mathbf{x}_{r,j}$ and $\mathbf{x}_{nr,j}$ denote the corresponding values of the selected markers of the single responder and non-responder respectively left out in the construction of the score. Now for each investigated γ_i the following function is calculated:

$$d_j(\mathbf{x}_{r,j}, \mathbf{x}_{nr,j};\gamma_i) = \begin{cases} 0 & \text{if } \hat{f}_j(\mathbf{x}_{r,j};\gamma_i) < \hat{f}_j(\mathbf{x}_{nr,j};\gamma_i) \\ 1 & \text{if } \hat{f}_j(\mathbf{x}_{r,j};\gamma_i) > \hat{f}_j(\mathbf{x}_{nr,j};\gamma_i) \\ 0.5 & \text{if } \hat{f}_j(\mathbf{x}_{r,j};\gamma_i) = \hat{f}_j(\mathbf{x}_{nr,j};\gamma_i) \end{cases} . \tag{4.5}$$

If no prediction score is selected from the data, $d_j(\mathbf{x}_{r,j}, \mathbf{x}_{nr,j};\gamma_i) = 0.5$. The values of $d_j(\mathbf{x}_{r,j}, \mathbf{x}_{nr,j};\gamma_i)$ for each γ_i, $i = 1, ..., s$, are now calculated for all $j = 1, ..., n^2$ training samples. Note that in our balanced scenario overall we use n^2 pairs of a single responder and non-responder as validation sample. For each γ_i we can calculate a "jackknife based" AUC_{γ_i}:

$$AUC_{\gamma_i} = \frac{1}{n^2} \sum_{j=1}^{n^2} d_j(\mathbf{x}_{r,j}, \mathbf{x}_{nr,j};\gamma_i). \tag{4.6}$$

This is the way the AUC would be calculated in the independent sample case (see e.g. Acion et al. (2006)). Finally we choose the selection boundary $\hat{\gamma}$ such that it maximizes AUC_{γ_i}:

$$\hat{\gamma} = \arg \max_{\gamma_i, i=1,...,s} \sum_{j=1}^{n^2} d_j(\mathbf{x}_{r,j}, \mathbf{x}_{nr,j};\gamma_i) \tag{4.7}$$

If more than one γ_i fulfills this jackknife criterion (4.7), the minimum of these γ_i values is chosen as final selection boundary.

We decided to search for the optimal γ_i instead of the optimal FDR because of the extremely longer runtime needed to search for the corresponding γ_i values in each training set. For the final selection boundary $\hat{\gamma}$ the FDR can be estimated with Storey's estimator (see formula (2.6)) in the total sample. However, this estimator may be biased.

Additionally the probability that at least one true null hypotheses is selected for future prediction

$$FWER_{\hat{\gamma}} = (1 - (1 - \hat{\gamma})^{m\pi_0}) \tag{4.8}$$

can be calculated. Despite the long runtime, for some scenarios we performed simulations of the procedure searching for the optimal FDR. The results are very similar.

Simulation results

To investigate the jackknife procedure we performed simulations for the scenarios assuming $m_e = 10$ and 60 among $m = 1000$ and 6000 hypotheses for $n = 50$ per group. For the scenarios with $m = 1000$ tested hypotheses we also investigated the procedure for $n = 100$.

$m_e = 10,\ m = 1000$

Figures 4.19 show the results of the jackknife procedure assuming 10 alternatives among 1000 hypotheses and fixing the sample size to $n = 50$ per group. Figure 4.19 (A) shows the "jackknife based" AUC_{γ_i} (see formula 4.6) as a function of γ_i ($i = 1, .., s$). The grey curves show AUC_{γ_i} of the 500 simulated samples and the black solid curve shows the mean value over the samples. For smaller values of γ_i on average larger jackknife based AUC values are achieved. Figure 4.19 (B) shows a boxplot of the final selection boundaries $\hat{\gamma}$ determined from (4.7). The jackknife procedure results in a median $\hat{\gamma}$ of 0.0021 (mean: 0.0052).

Figure 4.19 (C) shows boxplots of the estimated $\widehat{FDR}$ using Storey's estimator (see Formula (2.6)) for the different samples and the actual (true) FDR calculated from the samples. A boxplot of the asymptotic $FDR_{\hat{\gamma},\infty}$ calculated from the final selection boundaries $\hat{\gamma}$, assuming that π_0 and Δ are known can also be seen in Figure 4.19 (C). $FDR_{\hat{\gamma},\infty}$ is calculated as in formula (3.3) using

$$1 - \beta(\hat{\gamma}) = 1 - \Phi_{\sqrt{\frac{n}{2}}\Delta,1}(c_{1-\hat{\gamma}})$$

for the power. The jackknife procedure results in a mean $\widehat{FDR}$ of 0.243 (median: 0.19) which is only slightly smaller than the mean actual FDR of 0.254 (median: 0.2). Note that

the average asymptotic $FDR_{\hat{\gamma},\infty}$ is 0.256 (median: 0.19). Note also that if in the calculation of $\widehat{FDR}$ the estimated $\hat{\pi}_0$ is larger than 1, we set it to 1 (see Section 2.3.2). If no such correction of $\hat{\pi}_0$ is applied, the average $\widehat{FDR}$ is slightly larger that the actual FDR (mean: 0.265).

It remains the question how well the FDR is estimated when the threshold for the FDR is estimated itself from the re-sampling procedure and hence is a random variable. E.g. Genovese and Wasserman (2004) have given conditions for which asymptotically for large m (and large sample sizes) Storey's estimate for (random) thresholds is exceeding the true FDR with a given probability. To investigate the bias of the estimated FDR we calculated the difference between the estimated and the actual FDR (further on denoted by "bias"). Figure 4.19 (D) shows the boxplot of the differences which vary around 0.

A boxplot of the maxima of the jackknife based AUC_{γ_i} values, $AUC_{\hat{\gamma}} = \max_i AUC_{\gamma_i}$, over the simulated samples can be seen in Figure 4.19 (E). The average $AUC_{\hat{\gamma}}$ is 0.956. Figure (E) also shows a boxplot of the $AUC_{\hat{\gamma}}(\mathbf{x})$ values for the prediction of the outcome of a future patient (with marker values $\mathbf{x}$) calculated with formula (4.3) using the mean value μ_f and the variance σ_f^2 (see formula (4.2)) of each simulated sample. The average "future" $AUC_{\hat{\gamma}}(\mathbf{x})$ is 0.934. The theoretically best achievable $AUC_* = 0.965$ is shown as dotdashed horizontal line. Note that $AUC_{\hat{\gamma}}$ calculated from the data may be larger than $AUC_{\hat{\gamma}}(\mathbf{x})$ for prediction of a future patient and thus also larger than AUC_*.

Remember that the previous simulations in Section 4.6.1 (see Figure 4.4 (A)) showed that the largest average future AUC over the investigated FDR values (AUC_*^{sim}) occurs for a FDR of 0.15 and 0.2. Thus for this scenario, the selection boundaries $\hat{\gamma}$ found with the jackknife procedure lead to a median $\widehat{FDR}$ value which lies within the area where AUC_*^{sim} was found.

Figure 4.19 (F) shows a boxplot of the $FWER_{\hat{\gamma}}$ values (see formula (4.8)) assuming that π_0 is known. The mean value is 0.756 (median: 0.875). The actual FWER calculated from the 500 samples is 0.76.

Note that if more than one γ_i fulfills the jackknife criterion (4.7), we choose the smallest γ_i as selection boundary for future prediction. Since the γ_i values which achieve the same value lie within a small range, the results of the jackknife procedure when using the largest γ_i instead of the smallest are similar.

Increasing the sample size to $n = 100$ per group the jackknife procedure again performs good when assuming $m_e = 10$. Note that for $n = 100$ there are $n^2 = 10000$ possibilities of leaving out a pair of one responder and one non-responder respectively. A median $\hat{\gamma}$ of 0.0008 (mean: 0.0026) was found from the jackknife procedure which corresponds to a median actual FDR of 0.091 (mean: 0.1637) leading to a mean future $AUC_{\hat{\gamma}}(\mathbf{x})$ of 0.958, which is again only slightly smaller than AUC_*^{sim} of 0.961 found from the previous simulations (compare Figure 4.4 (B)). The average jackknife based $AUC_{\hat{\gamma}}$ is 0.965.

$m_e = 60$, $m = 1000$

Figures 4.20 show the results of the jackknife procedure assuming $m_e = 60$ alternatives among $m = 1000$ hypotheses and again applying a sample size of $n = 50$ per group. Figure 4.20 (A) shows that the jackknife based AUC_{γ_i} increases for increasing values of γ_i, however, the mean curve (black line) is flat for larger γ_i. A boxplot of the finally chosen $\hat{\gamma}$ values for the simulated samples is shown in Figure 4.20 (B). The median $\hat{\gamma}$ found from the jackknife procedure is 0.06 (mean: 0.097) which corresponds to a median estimated $\widehat{FDR}$ of 0.643 (mean: 0.615, Figure 4.20 (C)). The median actual FDR is 0.645 (mean: 0.604, Figure 4.20 (C)). Note that in this scenario the mean estimated $\widehat{FDR}$ is only slightly larger than the mean actual FDR, hence, the difference between both FDR values again varies around 0 (Figure 4.20 (D)), however, varying more than in the situation of $m_e = 10$. The median asymptotic $FDR_{\hat{\gamma},\infty}$ is 0.636 (mean: 0.613). Figure 4.20 (E) again shows boxplots of $AUC_{\hat{\gamma}}$ achieved for the final selection boundaries $\hat{\gamma}$ and the resulting future $AUC_{\hat{\gamma}}(\mathbf{x})$ for the prediction of the outcome of a future patient. The average $AUC_{\hat{\gamma}}$ is 0.868 as compared to a mean future $AUC_{\hat{\gamma}}(\mathbf{x})$ of 0.796, which is smaller than $AUC_*^{sim} = 0.813$ observed from the

previous simulations (compare Figure 4.6 (A)) and was achieved when controlling a FDR of 0.8 and 0.85. As expected in the situation of many effective marker with a rather small effect size the jackknife procedure works not as good as in the situation of a few alternatives with large effects. A smaller FDR is obtained from the jackknife procedure, however, because of the flat optimum with similar results over a wide range of FDR values, the performance in terms of the future $AUC_{\hat{\gamma}}(\mathbf{x})$ is still good. The median $FWER_{\hat{\gamma}}$ in this scenario is 1 (mean: 0.988, Figure 4.20 (F)). The actual FWER calculated from the 500 simulation steps is 0.984.

Increasing the sample size to $n = 100$, the jackknife procedure results in a median $\hat{\gamma}$ of 0.032 (mean: 0.044), which corresponds to a median actual FDR of 0.432 (mean: 0.405) leading to a mean $AUC_{\hat{\gamma}}(\mathbf{x})$ of 0.882. The mean $AUC_{\hat{\gamma}}(\mathbf{x})$ is only slightly smaller than AUC_*^{sim} of 0.887 which in the previous simulations occurred for a FDR between 0.45 and 0.55 (compare Figure 4.6 (B)). On average an $AUC_{\hat{\gamma}}$ of 0.957 is achieved. As expected, with an increased sample size of $n = 100$ per group the jackknife procedure performs better. The $\hat{\gamma}$ and FDR values found from the jackknife procedure lead to only slightly smaller performances in terms of future $AUC_{\hat{\gamma}}(\mathbf{x})$ values than found from our previous simulations.

$m_e = 10,\ m = 6000$

From the simulations in the previous sections it can be seen that when increasing the number of tested hypotheses to $m = 6000$ it is more difficult to find the effective markers. However, for $m_e = 10$ we may get a good performance of $AUC_*^{sim} = 0.918$ if the FDR is set to 0.25 (see Figure 4.7 (A)). Note that in order to get an overview of the scenario and to save runtime we only performed 100 simulation steps for the jackknife procedure. The jackknife procedure results in a median $\hat{\gamma}$ of 0.0005 (mean: 0.0011, see Figure 4.21 (B)) corresponding to a median estimated $\widehat{FDR}$ of 0.250 (mean: 0.286, Figure 4.21 (C)) leading to a mean future $AUC_{\hat{\gamma}}(\mathbf{x})$ of 0.910 (Figure 4.21 (E)) which is again only slightly smaller than AUC_*^{sim}. The average estimated $\widehat{FDR}$ is slightly smaller than the mean actual FDR of 0.298 (median: 0.226), the difference again varying around 0 (Figure 4.21 (D)). The mean asymptotic $FDR_{\hat{\gamma},\infty}$ is 0.314 (median: 0.260). The jackknife procedure again found

estimated $\widehat{FDR}$ values that lead to a good performance in terms of future $AUC_{\hat{\gamma}}(\mathbf{x})$ values. The mean jackknife based $AUC_{\hat{\gamma}}$ is 0.945 (Figure 4.21 (E)). Again the probability of at least one false positive is very large. The average $FWER_{\hat{\gamma}}$ is 0.787 (median: 0.933). The actual FWER calculated from the 100 simulation steps performed for $m = 6000$ is 0.75.

$m_e = 60,\ m = 6000$

Assuming $m_e = 60$ alternatives among the 6000 hypotheses resulted in a very poor performance of the selected prediction scores (see Figure 4.8 (A)). An AUC_*^{sim} of 0.669 was achieved for FDR= 0.9. The jackknife procedure results in a median selection boundary $\hat{\gamma}$ of 0.013 (mean: 0.034, see Figure 4.22 (B)) corresponding to a median estimated $\widehat{FDR}$ of 0.835 (mean: 0.812, see Figure 4.22 (C)). The median actual FDR is 0.817 (mean: 0.783). The mean estimated $\widehat{FDR}$ is again slightly larger than the mean actual FDR, however again the difference is varying around 0 (see Figure 4.22 (D)). Note that the median asymptotic $FDR_{\hat{\gamma},\infty}$ calculated from the $\hat{\gamma}$ values assuming that π_0 and Δ are known is 0.820 (mean: 0.801). The selection boundaries calculated with the jackknife procedure lead to a mean future $AUC_{\hat{\gamma}}(\mathbf{x})$ of 0.667 (see Figure 4.22 (E)). The mean jackknife based $AUC_{\hat{\gamma}}$ is 0.802. Only a very small AUC_*^{sim} in this scenario was obtained from the simulations, however the jackknife procedure leads to a mean $AUC_{\hat{\gamma}}(\mathbf{x})$ which is only slightly smaller than AUC_*^{sim}. Note that the average $FWER_{\hat{\gamma}}$ is 0.998 (median: 1) and the actual FWER calculated from the simulated samples is 0.98

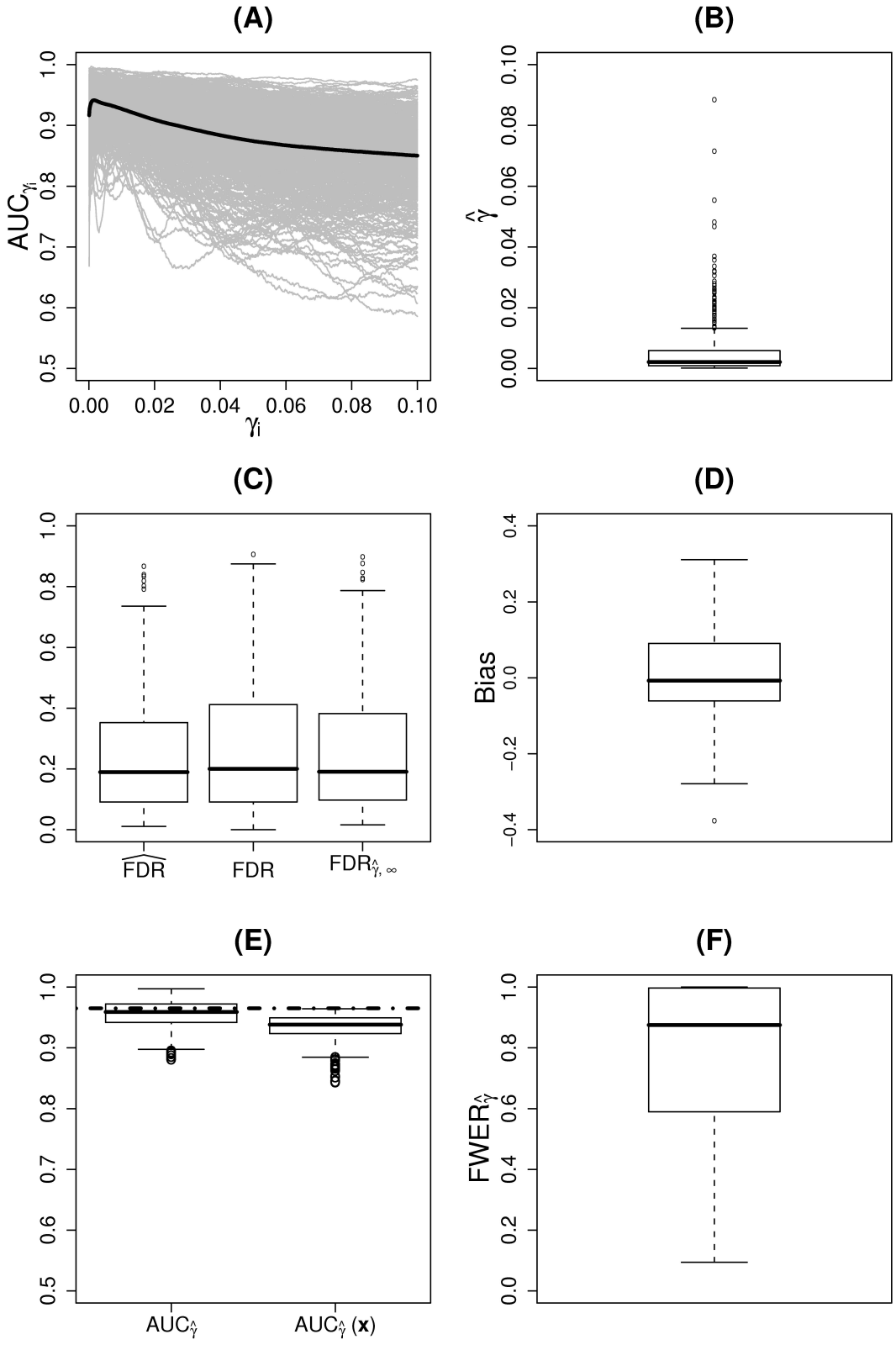

Figure 4.19: Simulation results (500 repetitions) of the Jackknife procedure applying the protected approach: AUC_{γ_i} as a function of γ_i (Figure (A)) as well as boxplots of the final selection boundary $\hat{\gamma}$ (B), the resulting estimated $\widehat{FDR}$, actual FDR and asymptotic $FDR_{\hat{\gamma},\infty}$ (C), the bias of the estimated FDR (D), the jackknife $AUC_{\hat{\gamma}}$ and the future $AUC_{\hat{\gamma}}(\mathbf{x})$ (E) and $FWER_{\hat{\gamma}}$ (F) for the situation of $m_e = 10$, $m = 1000$, $n = 50$.

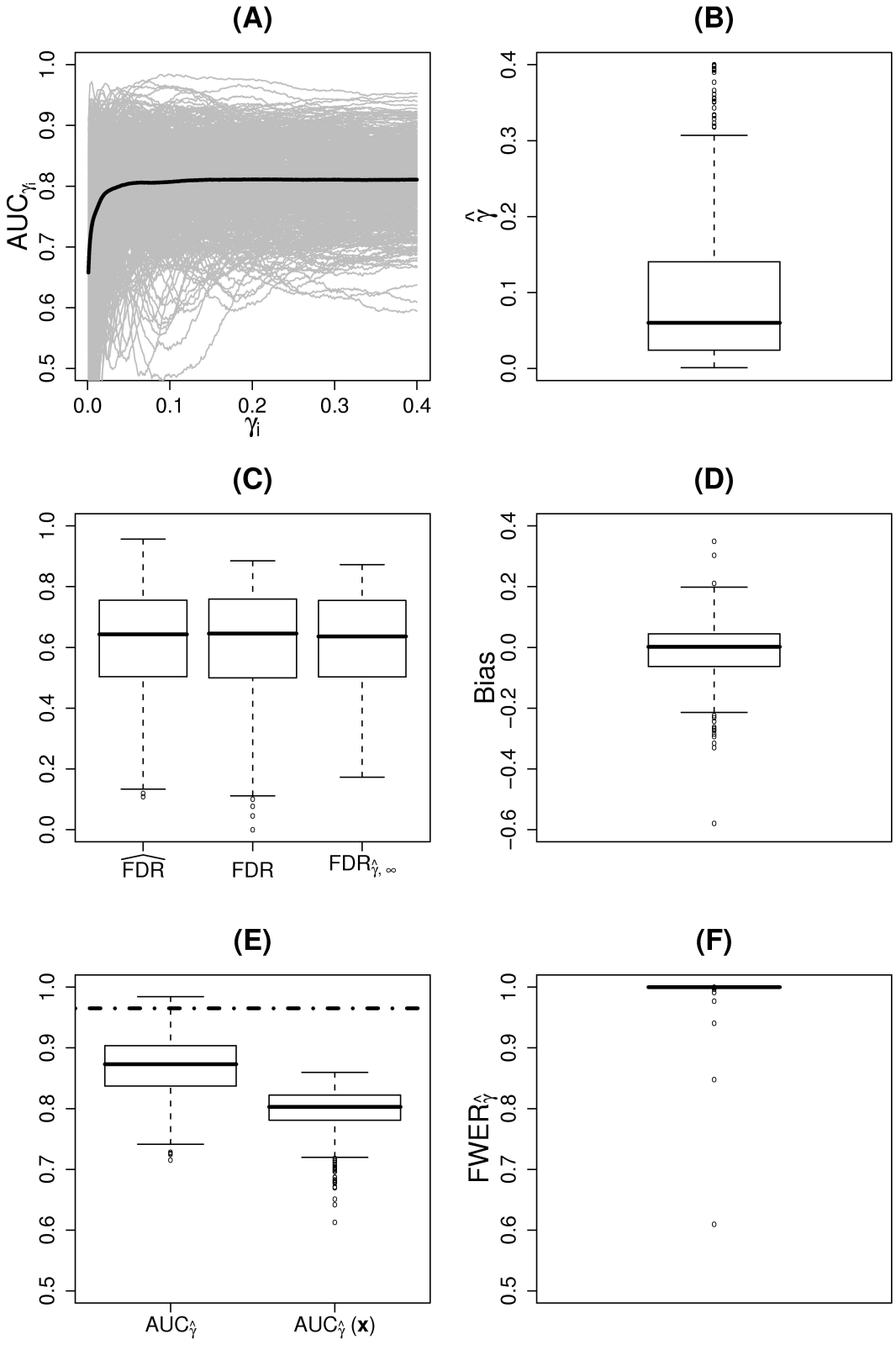

Figure 4.20: Simulation results (500 repetitions) of the Jackknife procedure applying the protected approach: AUC_{γ_i} as a function of γ_i (Figure (A)) as well as boxplots of the final selection boundary $\hat{\gamma}$ (B), the resulting estimated $\widehat{FDR}$, actual FDR and asymptotic $FDR_{\hat{\gamma},\infty}$ (C), the bias of the estimated FDR (D), the jackknife $AUC_{\hat{\gamma}}$ and the future $AUC_{\hat{\gamma}}(\mathbf{x})$ (E) and $FWER_{\hat{\gamma}}$ (F) for the situation of $m_e = 60$, $m = 1000$, $n = 50$.

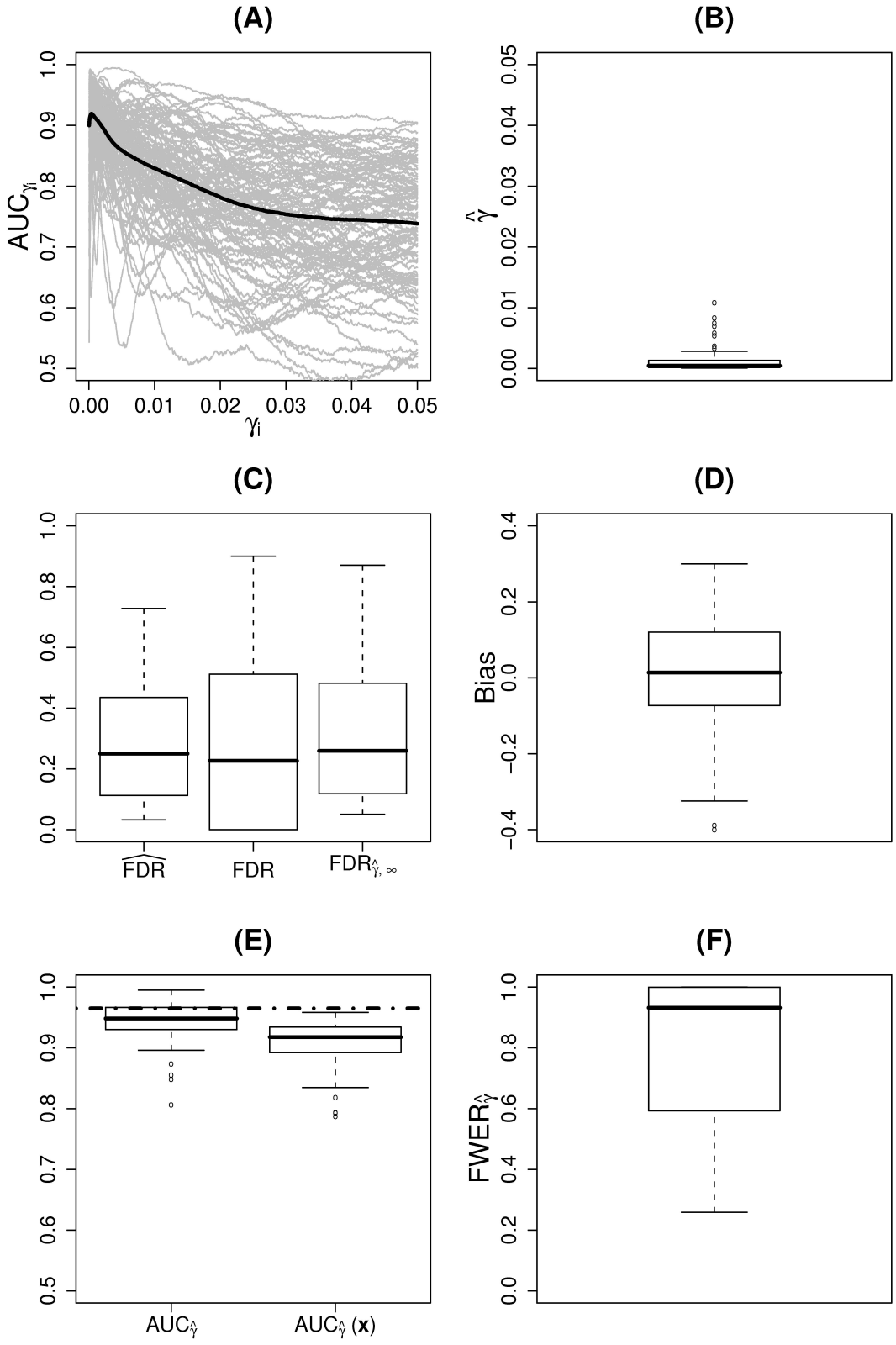

Figure 4.21: Simulation results (100 repetitions) of the Jackknife procedure applying the protected approach: AUC_{γ_i} as a function of γ_i (Figure (A)) as well as boxplots of the final selection boundary $\hat{\gamma}$ (B), the resulting estimated $\widehat{FDR}$, actual FDR and asymptotic $FDR_{\hat{\gamma},\infty}$ (C), the bias of the estimated FDR (D), the jackknife $AUC_{\hat{\gamma}}$ and the future $AUC_{\hat{\gamma}}(\mathbf{x})$ (E) and $FWER_{\hat{\gamma}}$ (F) for the situation of $m_e = 10$, $m = 6000$, $n = 50$.

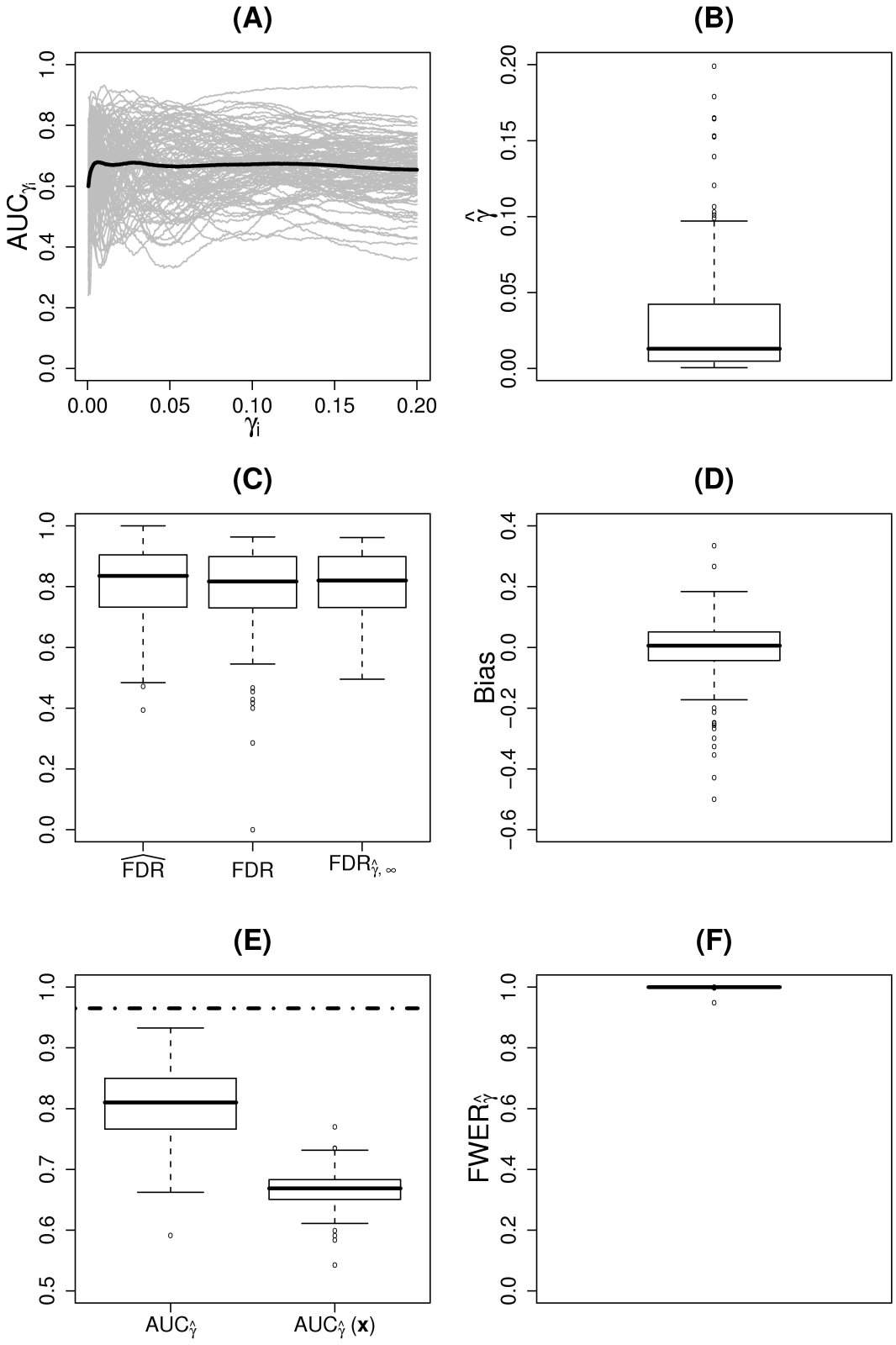

Figure 4.22: Simulation results (100 repetitions) of the Jackknife procedure applying the protected approach: AUC_{γ_i} as a function of γ_i (Figure (A)) as well as boxplots of the final selection boundary $\hat{\gamma}$ (B), the resulting estimated $\widehat{FDR}$, actual FDR and asymptotic $FDR_{\hat{\gamma},\infty}$ (C), the bias of the estimated FDR (D), the jackknife $AUC_{\hat{\gamma}}$ and the future $AUC_{\hat{\gamma}}(\mathbf{x})$ (E) and $FWER_{\hat{\gamma}}$ (F) for the situation of $m_e = 60$, $m = 6000$, $n = 50$.

4.7.2 Jackknife for the optimistic approach

The jackknife procedure described for the protected method can be easily applied to the unprotected method selecting the best k markers for future prediction by searching within a grid of values of k_i instead of the γ_i values. The jackknife criterion is given by replacing γ_i by k_i in formulas (4.7) and (4.5). The procedure itself remains the same.

m=1000

Figures 4.23 and 4.24 shows the results of the jackknife procedure for selection of best k if $m_e = 10$ and $m_e = 60$ alternatives (effective markers) are assumed among 1000 hypotheses. The sample size per group is fixed to $n = 50$ per group. For $m_e = 10$ the jackknife procedure results in a median $\hat{k}$ of 11 (mean: 14.58, Figure 4.23 (A)), which achieves the largest jackknife based $AUC_{\hat{k}}$ and leads to a mean $AUC_{\hat{k}}(\mathbf{x})$ for future prediction of 0.939 (Figure 4.23 (B)). The average $AUC_{\hat{k}}$ is 0.959 (Figure 4.23 (B)). A median number of 9 (mean: 8.8) out of the 10 alternatives are selected for future prediction. Remembering the simulations in Section 4.6.3 (see Figure 4.13 (A)), $AUC_*^{sim} = 0.942$ was found for $k = 10$. The jackknife procedure results in $\hat{k}$ values which in average lead to only a slightly smaller future performance.

For the situation of $m_e = 60$ alternatives among 1000 hypotheses a median $\hat{k}$ of 125 (mean: 212.99) was determined from the jackknife procedure (see Figures 4.24), which leads to a mean $AUC_{\hat{k}}(\mathbf{x})$ of 0.809 which is slightly smaller than AUC_*^{sim} found in the simulations. Note that the optimum of the average AUC depending on k was very flat. Virtually the same performance of in average 0.814 was achieved for $k = 150$ to 600 (see Figure 4.15 (A)). Note that the average $AUC_{\hat{k}}$ is 0.866 which is again larger than the average $AUC_{\hat{k}}(\mathbf{x})$ for prediction of a future patient. A median number of 38 (mean: 39.8) out of the 60 alternatives are selected for future prediction.

m=6000

Increasing the number of tested markers to 6000 the jackknife procedure for $m_e = 10$ again performs well (see Figures 4.25). The jackknife procedure found a median $\hat{k}$ of 9 (mean: 13.04) which results in an average $AUC_{\hat{k}}(\mathbf{x})$ of 0.915 which is again only slightly smaller than AUC_*^{sim} of 0.919 found in our simulations (compare Figure 4.16 (A)). A median number of 8 (mean: 7.7) out of the 10 alternatives are selected for future prediction. The mean $AUC_{\hat{k}}$ is 0.946. Although the overall results in terms of future AUC values assuming $m_e = 60$ alternatives among 6000 hypotheses are poor over the investigated values of k (compare Figures 4.17) and the optimum is very flat, the jackknife procedure resulted in a median $\hat{k} = 148$ (mean: 242.05, see Figures 4.26) leading to prediction scores with an average $AUC_{\hat{k}}(\mathbf{x})$ of 0.671 which is again only slightly smaller than AUC_*^{sim} of 0.673 achieved for k between 200 and 300. The average jackknife $AUC_{\hat{k}}$ is 0.769. A median number of 21.5 (mean: 23.4) out of the 60 alternatives is selected for future prediction.

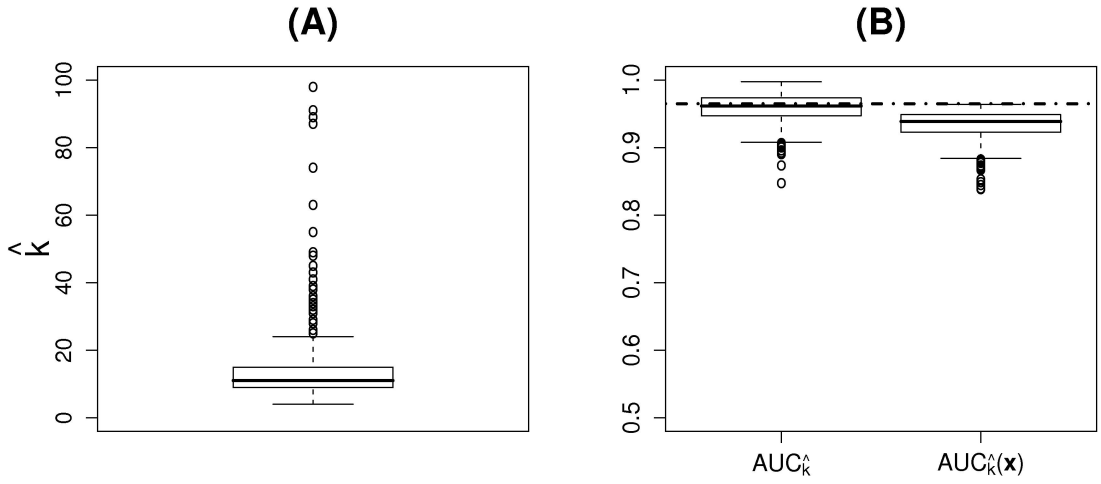

Figure 4.23: Simulation results (500 repetitions) for the Jackknife procedure applying the optimistic approach by selection of the best k markers: Boxplots of the $\hat{k}$ resulting from the jackknife procedure (Figure (A)) as well as $AUC_{\hat{k}}$ and $AUC_{\hat{k}}(\mathbf{x})$ (Figure (B)) for the situation of $m_e = 10$, $m = 1000$, $n = 50$.

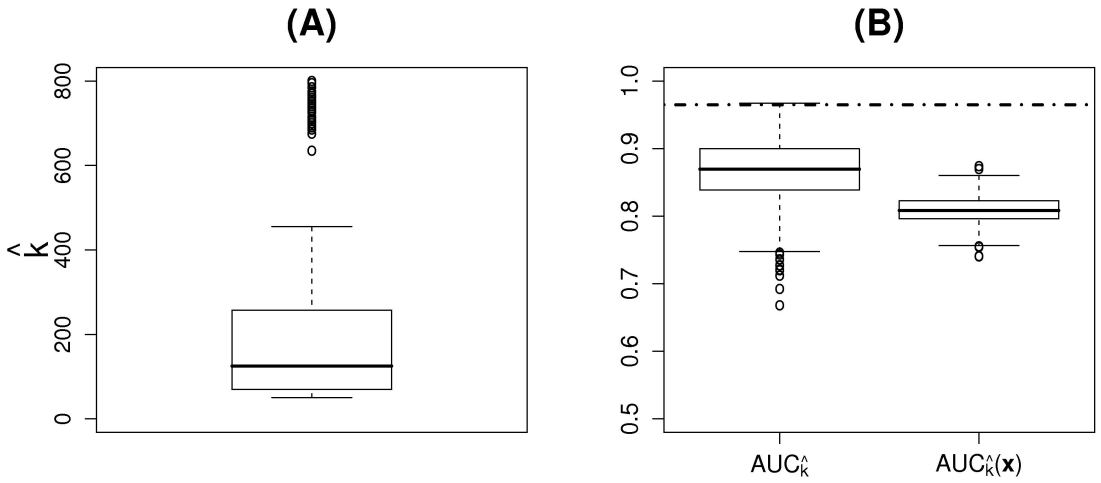

Figure 4.24: Simulation results (500 repetitions) for the Jackknife procedure applying the optimistic approach by selection of the best k markers: Boxplots of the $\hat{k}$ resulting from the jackknife procedure (Figure (A)) as well as $AUC_{\hat{k}}$ and $AUC_{\hat{k}}(\mathbf{x})$ (Figure (B)) for the situation of $m_e = 60$, $m = 1000$, $n = 50$.

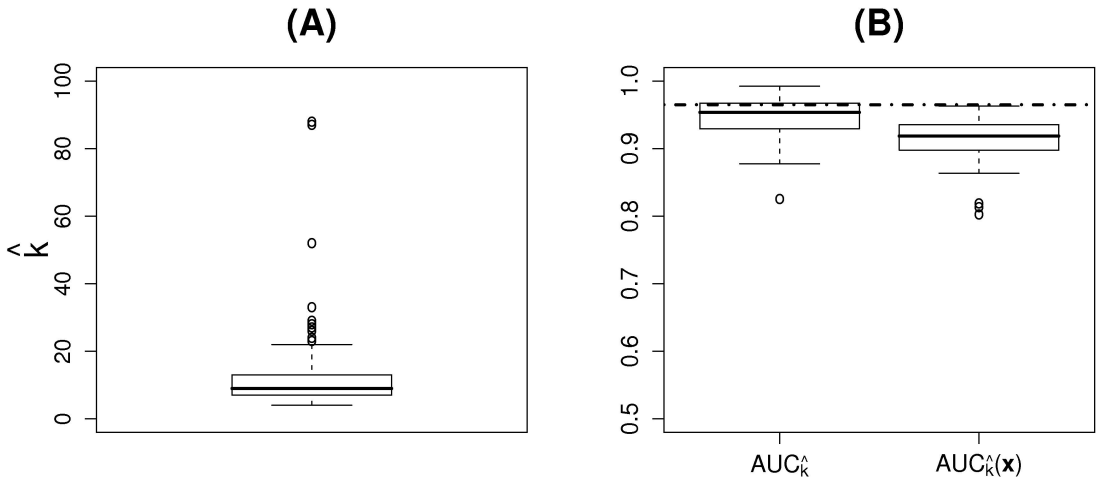

Figure 4.25: Simulation results (100 repetitions) for the Jackknife procedure applying the optimistic approach by selection of the best k markers: Boxplots of the $\hat{k}$ resulting from the jackknife procedure (Figure (A)) as well as $AUC_{\hat{k}}$ and $AUC_{\hat{k}}(\mathbf{x})$ (Figure (B)) for the situation of $m_e = 10$, $m = 6000$, $n = 50$.

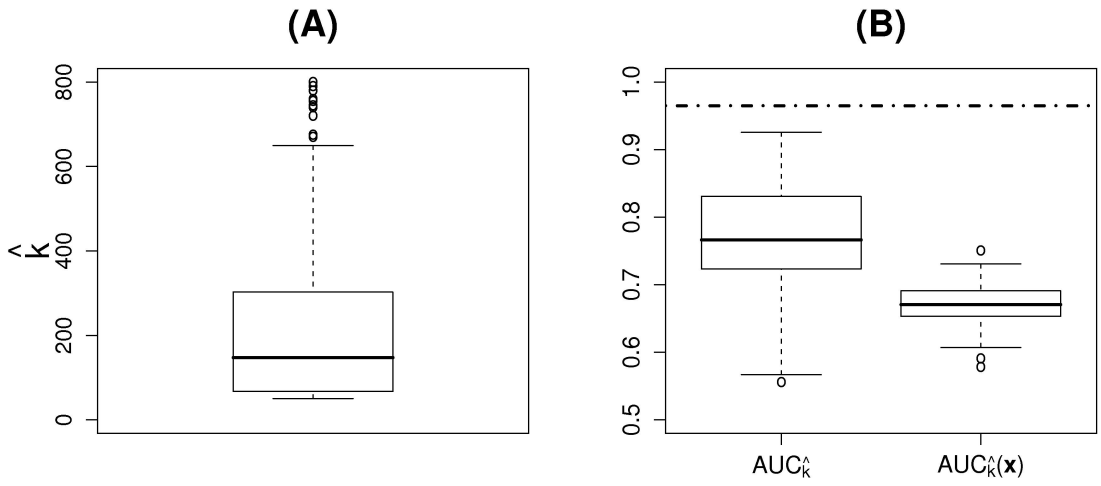

Figure 4.26: Simulation results (100 repetitions) for the Jackknife procedure applying the optimistic approach by selection of the best k markers: Boxplots of the $\hat{k}$ resulting from the jackknife procedure (Figure (A)) as well as $AUC_{\hat{k}}$ and $AUC_{\hat{k}}(\mathbf{x})$ (Figure (B)) for the situation of $m_e = 60$, $m = 6000$, $n = 50$.

4.7.3 Jackknife under the global null hypothesis

When using the protected procedure based on a multiple test controlling the FDR, the probability to end with useless prediction scores is targeted at the pre-chosen FDR. However, if we expect a large number of effective markers with rather small effects choosing a large FDR would be superior. Under the global null hypothesis the modified jackknife procedure to select decision boundaries which produce "best" cross validated ROC-curve will lead to select scores, however, they are useless for future patients.

Figures 4.27 and 4.28 show the results for the jackknife procedures under the global null hypothesis of no effective marker at all for selection using the protected approach assuming a sample size of $n = 50$ and $n = 100$ per group respectively if a overall number of 1000 hypotheses are tested. The average curve of the jackknife based AUC_{γ_i} over the simulated samples as a function of γ_i is nearly a horizontal line at $AUC_{\gamma_i} = 0.5$, indicating that no prediction score can be selected (Figures (A)). However, the jackknife procedure is searching for the selection boundary which maximizes this jackknife based AUC_{γ_i} so that in most cases the jackknife procedure will choose a prediction score in case of the global null (see the boxplots

for the jackknife based $\hat{\gamma}$: Figures 4.27 and 4.28 (B)). Only in a few simulation steps no prediction score is calculated (3.2% of the 500 steps if $n = 50$ and 2.2% if $n = 100$) under the global null. Note that in the scenarios simulated if the alternative holds (as considered in the previous sections), in all simulated samples a prediction score was selected from the data. A median $\hat{\gamma}$ of 0.0115 (mean: 0.0336) is found as final selection boundary if $n = 50$ and of 0.008 (mean: 0.0234) if n was set to 100 per group.

Figure 4.29 shows the results of the jackknife procedure testing 6000 hypotheses fixing $n = 50$ per group. A median $\hat{\gamma}$ of 0.0016 (mean: 0.0029, see Figure 4.29 (B)) is found as final selection boundary. Again the average AUC_{γ_i}-curve is nearly horizontal at 0.5 (Figure 4.29 (A)). No prediction score was calculated in 6% of the 100 simulation steps.

The results when applying the optimistic approach are similar. Again the mean curve of the jackknife based AUC_{k_i} over the simulated samples for the investigated k_i is nearly a horizontal line at AUC= 0.5 (see Figure 4.30 (A) if $m = 1000$ and 4.31 (A) if $m = 6000$). However, the median jackknife based $\hat{k}$ chosen when the global null hypotheses is true is 7 (mean: 18.67, Figure 4.30 (B)) if $m = 1000$ and 5.5 (mean: 21.1, Figure 4.31 (B)) if $m = 6000$. In both scenarios n is set to 50.

To get a hint that the global null hypotheses is true it would be useful to look at the quality obtained by cross validated ROC-curves itself, e.g., in terms of the jackknife based $AUC_{\hat{\gamma}}$ achieved with the final selection boundaries $\hat{\gamma}$. Figure 4.32 shows the distribution of the jackknife based $AUC_{\hat{\gamma}}$ of the simulated samples for selection applying the protected approach for the situations of $m_e = 10$ (first row) and 60 (second row) among 1000 tested hypotheses as well as under the global null hypotheses (third row) fixing the sample size to $n = 50$ (fist column) and $n = 100$ (second column). The histograms show that in case of $m_e > 0$ $AUC_{\hat{\gamma}}$ is generally larger than 0.8 whereas in case of the global null $AUC_{\hat{\gamma}}$ is generally below 0.8. Thus, in this situation one criterion could be the following: if the $AUC_{\hat{\gamma}}$ resulting from the jackknife procedure is smaller than, e.g. 0.8 then it seems to be preferable not to

construct any score at all. Applying a sample size of $n = 100$ per group there is no overlap between the distributions under the alternative and under the global null hypothesis in the simulated samples. Expecting $m_e = 60$ and fixing $n = 50$ per group in 90.8% of the simulated samples $AUC_{\hat{\gamma}}$ is larger than 0.8, under the global null, in only 2.6% of the simulated samples.

Increasing the number of tested hypotheses to 6000 in the situation of $m_e = 10$ again $AUC_{\hat{\gamma}}$ is always larger than 0.8. However, when expecting a larger number of alternatives with rather small effect sizes the distribution of $AUC_{\hat{\gamma}}$ overlaps the distribution under the global null on a larger area. Deciding to construct a score if the $AUC_{\hat{\gamma}}$ is larger than 0.8 would lead to a false negative decision in 55% of the simulated samples. Decreasing the boundary to 0.7 in only 8% no prediction score would be constructed. However, under the global null hypotheses in 29% of the simulated samples $AUC_{\hat{\gamma}}$ is above 0.7. Note that the sample size in these scenarios is set to $n = 50$ per group.

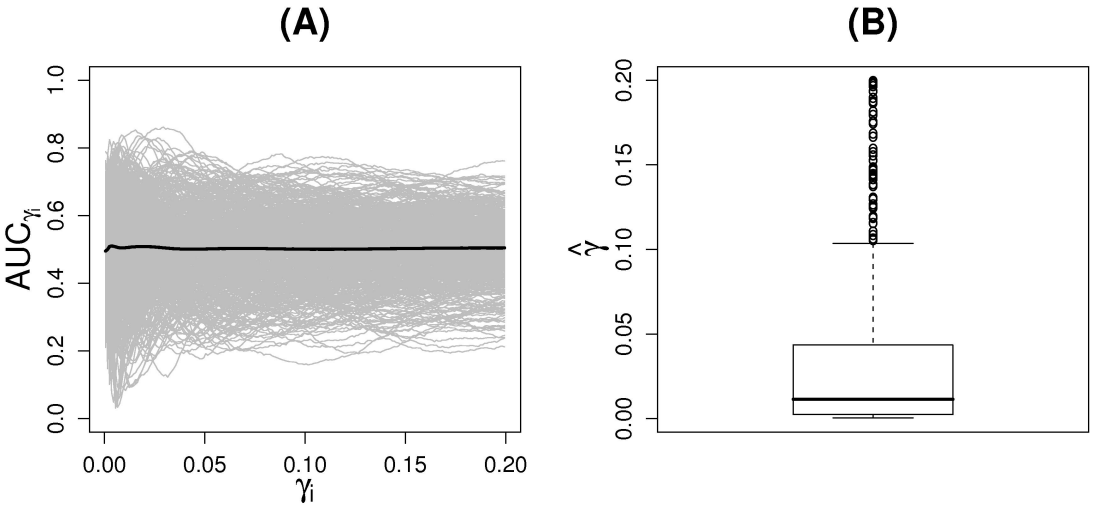

Figure 4.27: Simulation results (500 repetitions) for the Jackknife procedure applying the protected approach by selection using the FDR under the global Null: The jackknife based AUC_{γ_i} as a function of the selection boundaries γ_i (Figure (A)) as well as the boxplot of the final selection boundary $\hat{\gamma}$ resulting from the jackknife procedure (Figure (B)) for the situation of $m = 1000$, $n = 50$.

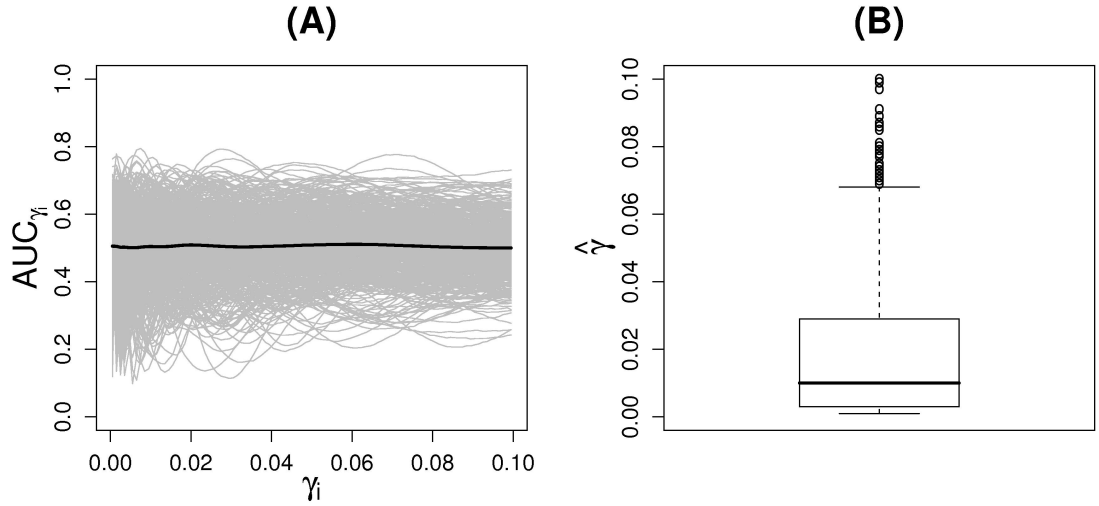

Figure 4.28: Simulation results (500 repetitions) for the Jackknife procedure applying the protected approach by selection using the FDR under the global Null: The jackknife based AUC_{γ_i} as a function of the selection boundaries γ_i (Figure (A)) as well as the boxplot of the final selection boundary $\hat{\gamma}$ resulting from the jackknife procedure (Figure (B)) for the situation of $m = 1000$, $n = 100$.

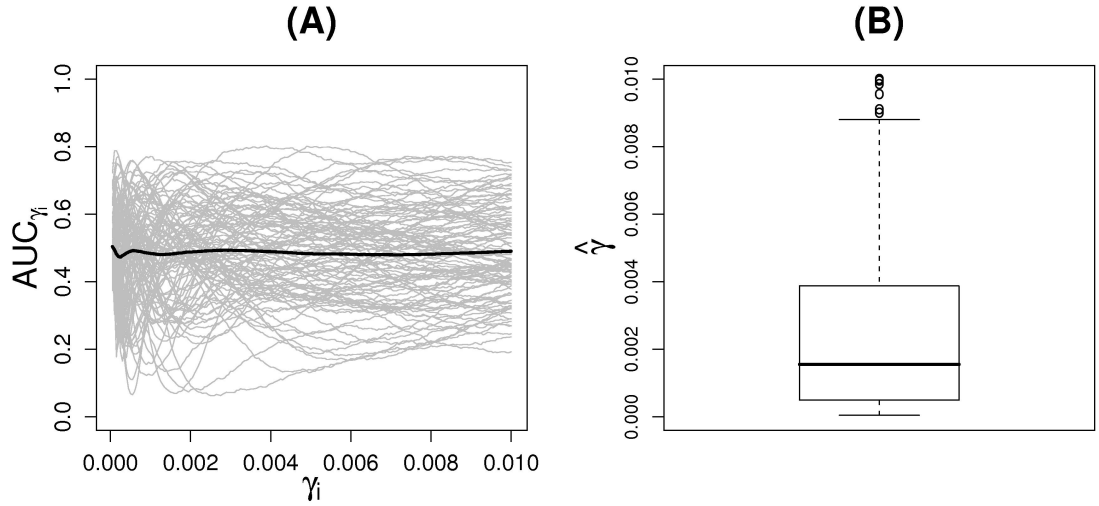

Figure 4.29: Simulation results (100 repetitions) for the Jackknife procedure applying the protected approach by selection using the FDR under the global Null: The jackknife based AUC_{γ_i} as a function of the selection boundaries γ_i (Figure (A)) as well as the boxplot of the final selection boundary $\hat{\gamma}$ resulting from the jackknife procedure (Figure (B)) for the situation of $m = 6000$, $n = 50$.

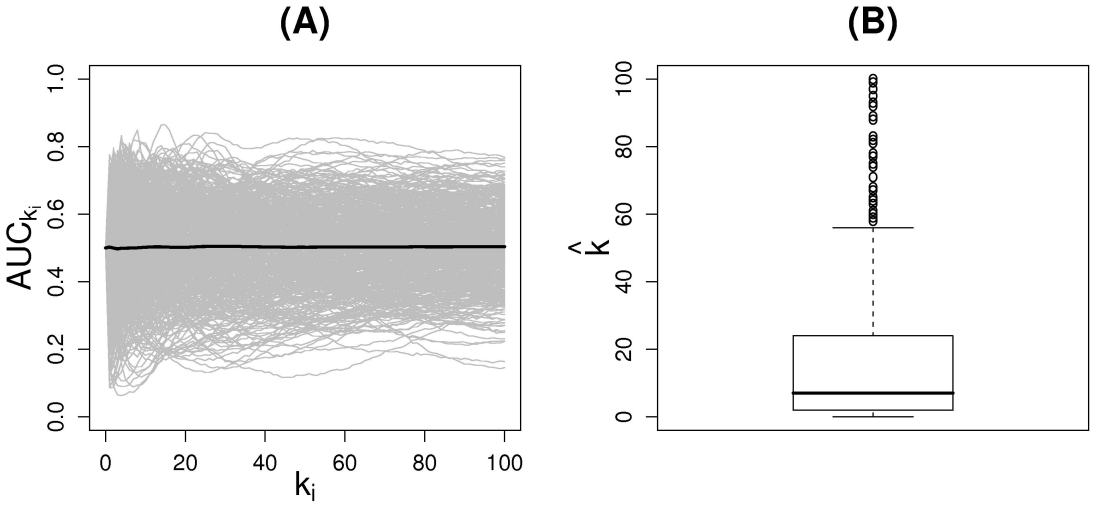

Figure 4.30: Simulation results (500 repetitions) for the Jackknife procedure applying the optimistic approach by selection of the best k markers under the global Null: The jackknife based AUC_{k_i} as a function of the selection boundaries k_i (Figure (A)) as well as the boxplot of the final selection boundary $\hat{k}$ resulting from the jackknife procedure (Figure (B)) for the situation of $m = 1000$, $n = 50$.

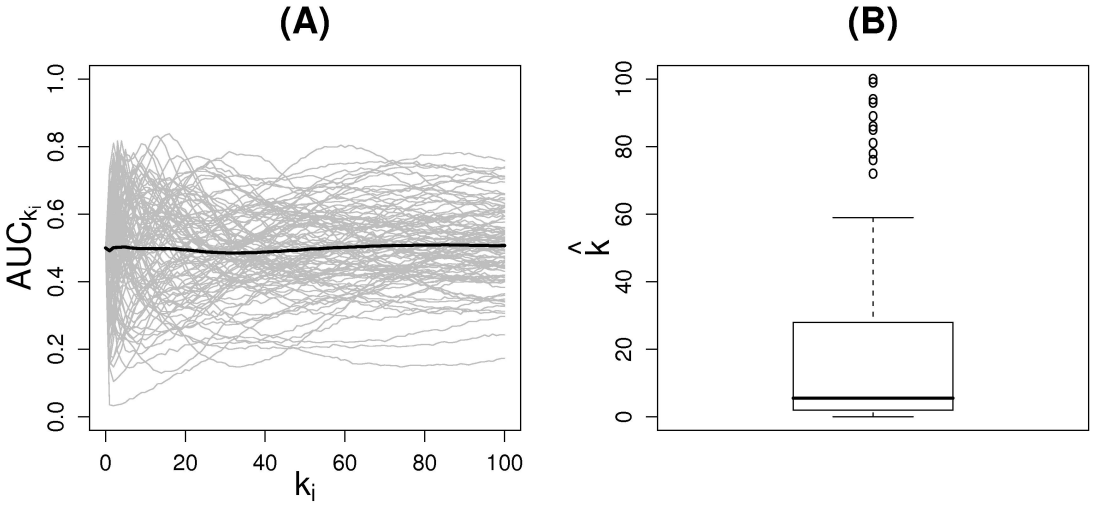

Figure 4.31: Simulation results (100 repetitions) for the Jackknife procedure applying the optimistic approach by selection of the best k markers under the global Null: The jackknife based AUC_{k_i} as a function of the selection boundaries k_i (Figure (A)) as well as the boxplot of the final selection boundary $\hat{k}$ resulting from the jackknife procedure (Figure (B)) for the situation of $m = 6000$, $n = 50$.

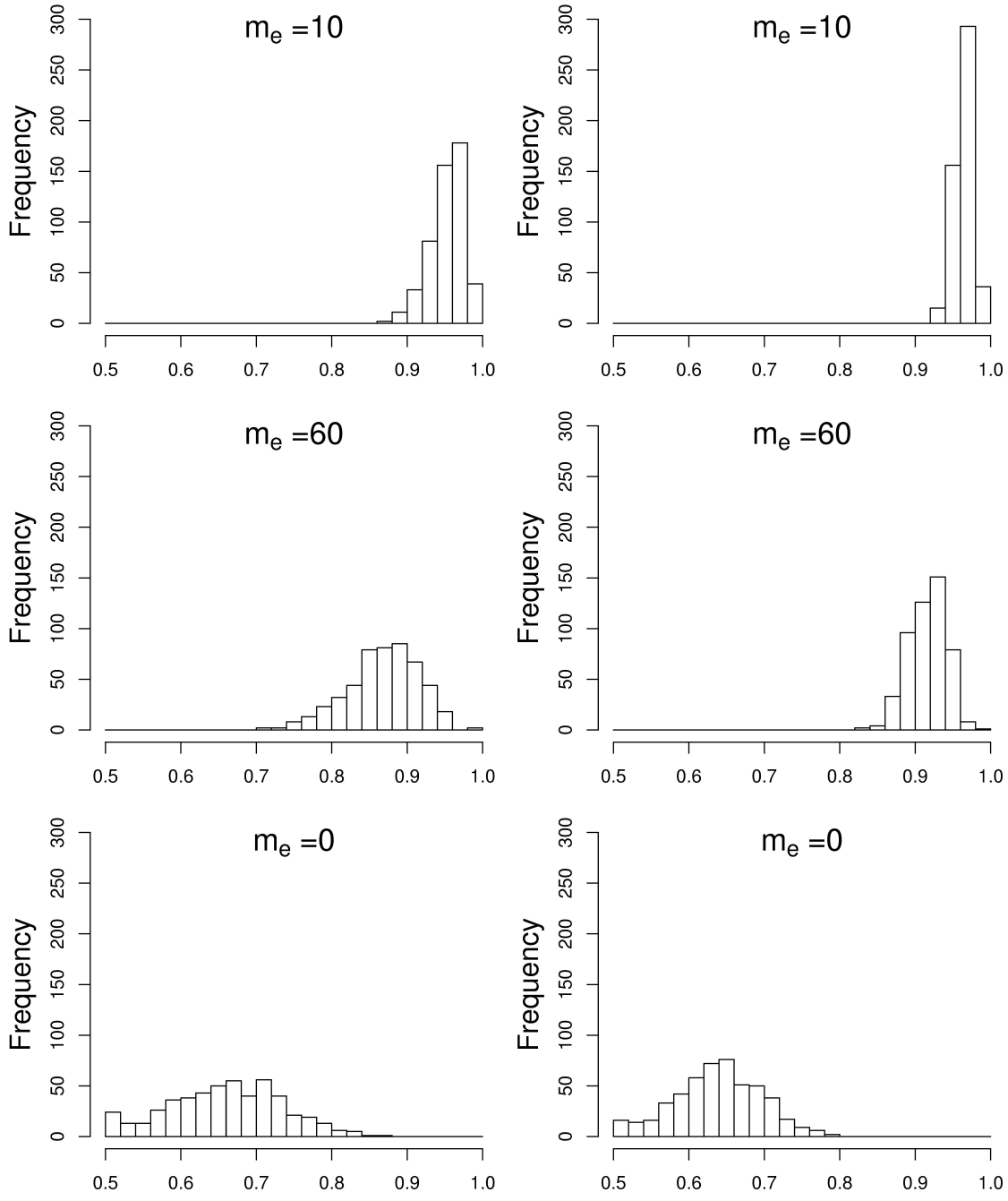

Figure 4.32: Distribution of the jackknife AUC for selection using FDR (500 simulation steps): $m_e = 10$, 60 or 0 among $m = 1000$ hypotheses. The sample size was set to $n = 50$ per group (first column) and $n = 100$ (second column).

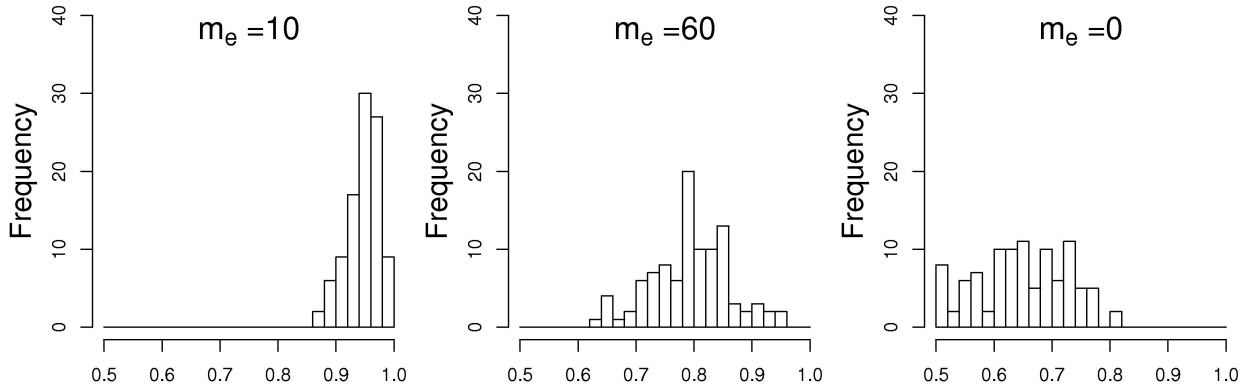

Figure 4.33: Distribution of the jackknife AUC for selection using FDR (100 simulation steps): $m_e = 10$, 60 or 0 among $m = 6000$ hypotheses. The sample size was set to $n = 50$ per group.

4.8 Variable Selection expecting a small AUC_*

In the last sections we assumed that the optimal linear prediction score of future patients, if known, would lead to a ROC-curve crossing through the benchmark point where sensitivity and specificity are 0.9, which corresponds to a theoretically achievable AUC_* of 0.965 indicating a very good discrimination between responders and non-responders. However, in medical research such AUC values may not be achieved. To investigate a more realistic scenario, we also simulated the situation where the AUC_* is assumed to be 0.8, which is, e.g., achievable in predicting hospital mortality from scores based on a set of variable measured in patients at admission to an intensive care unit (and constructed in large training samples, see Metnitz et al. (2005a) and Metnitz et al. (2005b)). We only investigated simulations for the protected approach using a multiple test controlling the FDR since we expect similar tendencies for the other selection method by selecting the best k markers. To achieve a future performance of $AUC_* = 0.8$ the benchmark point is at $v = 1 - w = 0.724$ (assuming a ROC crossing through a point, where sensitivity and specificity are the same). The minimal Δ required to obtain this ROC is 0.38 assuming $m_e = 10$ alternatives and 0.15 for $m_e = 60$. Figure 4.34 shows the minimal required Δ for different values for AUC_*. Results assuming 10 (dashed line) and 60 (solid line) effective markers are shown. The results for $AUC_* = 0.8$ and 0.965 (as assumed in the previous sections) are marked.

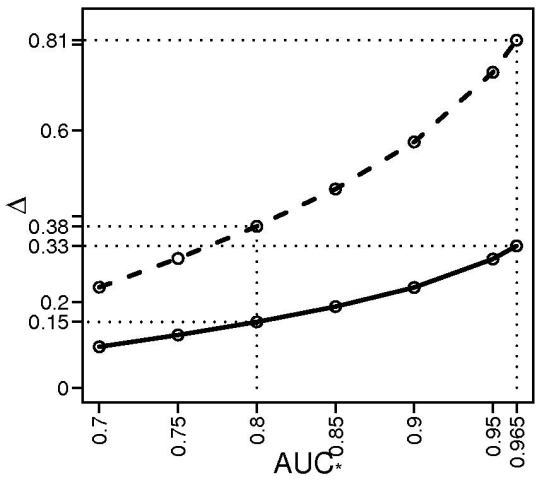

Figure 4.34: Minimal required effect size Δ as a function of AUC_* assuming $m_e = 10$ (dashed line) and 60 (solid line)

4.8.1 Simulation results

We again performed simulations (10000 simulation steps) assuming $m_e = 10$ and 60 alternatives among $m = 1000$ and 6000 hypotheses. The sample size per group is set to $n = 50$, 100 and 500.

m=1000

Figure 4.35 shows the resulting AUC values expecting $m_e = 10$ among 1000 hypotheses fixing the sample size to $n = 50$ (Figure 4.35 (A)), 100 (Figure 4.35 (B)) and 500 (Figure 4.35 (C)) per group. Figure 4.35 (A) shows that applying a small sample size per group we will not be able to find good prediction scores whatever FDR is used for selection. The average AUC values conditional that at least one marker is selected for future prediction vary around 0.6 over the whole investigated grid of FDRs. Note that if a FDR of 0.05 is chosen, in only 24.7% of the simulated samples a prediction score is selected. This percentage increases to 99% if the FDR is targeted with 0.95. Larger sample sizes are needed to detect the alternatives with their only small effect sizes required to achieve $AUC_* = 0.8$. However, when doubling the sample size to $n = 100$ per group (Figure 4.35 (B)) AUC_*^{sim} only increases to

0.676 and occurs for FDR= 0.5. Fixing the sample size to $n = 500$ leads to prediction scores with $AUC_*^{sim} = 0.795$ for a FDR of 0.005 and 0.01 which is only slightly smaller than AUC_* of 0.8 (Figure 4.35 (C)).

Increasing the number of effective markers to $m_e = 60$ hypotheses (Figures 4.36) again a very large sample size is needed to achieve good prediction scores. However, fixing the sample size to $n = 500$ per group AUC_*^{sim} is only 0.720 for a FDR between 0.4 and 0.5. For $n = 50$ the average AUC values do not exceed 0.6 over the whole range of investigated FDR values. For $n = 100$ per group AUC_*^{sim} is 0.614 for FDR= 0.9.

m=6000

If the effective markers are searched among 6000 hypotheses the situation gets extremely worse if smaller sample sizes are considered (Figures 4.37 and 4.38). Increasing the sample size to $n = 500$ per group helps if only a small number of alternatives with large effect sizes is assumed. Whereas for $m_e = 10$ (Figure 4.37 (C)) an AUC_*^{sim} of 0.793 can be obtained if the FDR is set to 0.05, for $m_e = 60$, AUC_*^{sim} is only 0.657 for a FDR of 0.55 and 0.6 (Figure 4.38 (C)). Thus the conclusion is that if there is only a only moderate true discrimination between responders and non-responders very large sample sizes are required to get good prediction scores. Studies with small sizes will mostly produce useless prediction scores.

4.8.2 Jackknife procedure

m=1000, m_e=10

To investigate the jackknife procedure for the situation of a smaller AUC_*, we first investigate the scenario assuming $m_e = 10$ among $m = 1000$ for a sample size of $n = 50$ and $n = 100$ per group (500 simulation steps). Figures 4.39 show the results for $n = 50$. Figure 4.39 (A) shows the jackknife based AUC_{γ_i} values as a function of the selection boundaries γ_i. The mean curve (solid curve) over the simulated samples is flat over the investigated γ_i values.

The jackknife procedure results in a median final selection boundary $\hat{\gamma}$ of 0.0125 (mean: 0.0424, Figure 4.39 (B)) which corresponds to a median estimated $\widehat{FDR}$ of 0.763 (mean: 0.713, Figure (C)). The mean estimated $\widehat{FDR}$ is slightly larger than the mean actual FDR of 0.711 (median: 0.813). The difference between the estimated and the actual FDR is again varying around 0 (see Figure 4.39 (D)). Note that the median asymptotic $FDR_{\hat{\gamma},\infty}$ is 0.774 (mean: 0.734). The $\hat{\gamma}$ values found by the jackknife procedure lead to prediction scores with a mean future $AUC_{\hat{\gamma}}(\mathbf{x})$ of 0.598 (Figure 4.39 (E)). The average $AUC_{\hat{\gamma}}$ is 0.741. However, the average proportion of at least one false positive ($FWER_{\hat{\gamma}}$) is 0.910 (median: 1, Figure 4.39 (F)). The actual FWER calculated from the 500 simulation steps is 0.92. The similar performance in terms of future AUC values over the whole range of investigated FDR values with no explicit optimum clearly indicates no good condition to use the jackknife procedure.

Increasing the sample size to $n = 100$ the jackknife procedure found a median final selection boundary $\hat{\gamma}$ of 0.009 (mean: 0.022) corresponding to a median actual FDR of 0.613 (mean: 0.555) leading to prediction scores with an average $AUC_{\hat{\gamma}}(\mathbf{x})$ of 0.665. Remember that the AUC_*^{sim} found from the simulation was 0.676 for FDR= 0.5. Note that the mean jackknife $AUC_{\hat{\gamma}}$ is 0.760.

m=1000, m_e=60

Assuming $m_e = 60$ alternatives among 1000 tested hypotheses, the jackknife procedure results in a median $\hat{\gamma}$ of 0.039 (mean: 0.139, see Figure 4.40 (B)) which corresponds to a mean estimated $\widehat{FDR}$ of 0.815 (median: 0.852, Figure 4.40 (C)). The mean actual FDR of 0.779 (median: 0.833) is slightly smaller. The difference between both FDR values is again varying around 0 (Figure 4.40 (D)). The mean asymptotic $FDR_{\hat{\gamma},\infty}$ is 0.781 (median: 0.797). Figure 4.40 (E) shows that the selection boundaries selected by the jackknife procedure lead to prediction scores with an average future $AUC_{\hat{\gamma}}(\mathbf{x})$ of 0.551. Note that AUC_*^{sim} found in the simulations is 0.583 when controlling a FDR of 0.90. The average $AUC_{\hat{\gamma}}$ is 0.701. Note that $FWER_{\hat{\gamma}}$ is again very large with a mean value of 0.951 (median: 1) and the actual FWER calculated from the simulated sample is 0.962.

Global null hypothesis

Figure 4.41 describes the problem if the global null hypotheses is true. The Figures show the distributions of the $AUC_{\hat{\gamma}}$ values expecting $m_e = 10$, 60 and 0 alternatives assuming $AUC_* = 0.8$ if the alternative holds. For $m_e = 10$ the shift of the distribution to the left is about one standard deviation (a mean value of 0.741 compared to a mean value of 0.660 under the global null). Fixing the boundary for $AUC_{\hat{\gamma}}$ to e.g. 0.7 for the basic decision to construct a prediction score or not, in 26.6% of the simulated samples no prediction score would be calculated from the samples even if in reality a prediction score exists. In case of the global null hypotheses, for 32.6% of the simulated samples $AUC_{\hat{\gamma}} \geq 0.7$.
The shift between the distributions expecting $m_e = 60$ and the global null is only about one half standard deviation (a mean value of 0.701 as compared to 0.660). Thus in case of a moderate true discrimination between responders and non-responders ($AUC_* = 0.8$) and a large number of effective markers with small effect sizes inducing this discrimination it may become a formidable task to distinguish whether in reality the alternative or the global null holds. However, the small jackknife $AUC_{\hat{\gamma}}$ may also indicate that the effect sizes under the alternative are to small for the given sample size.

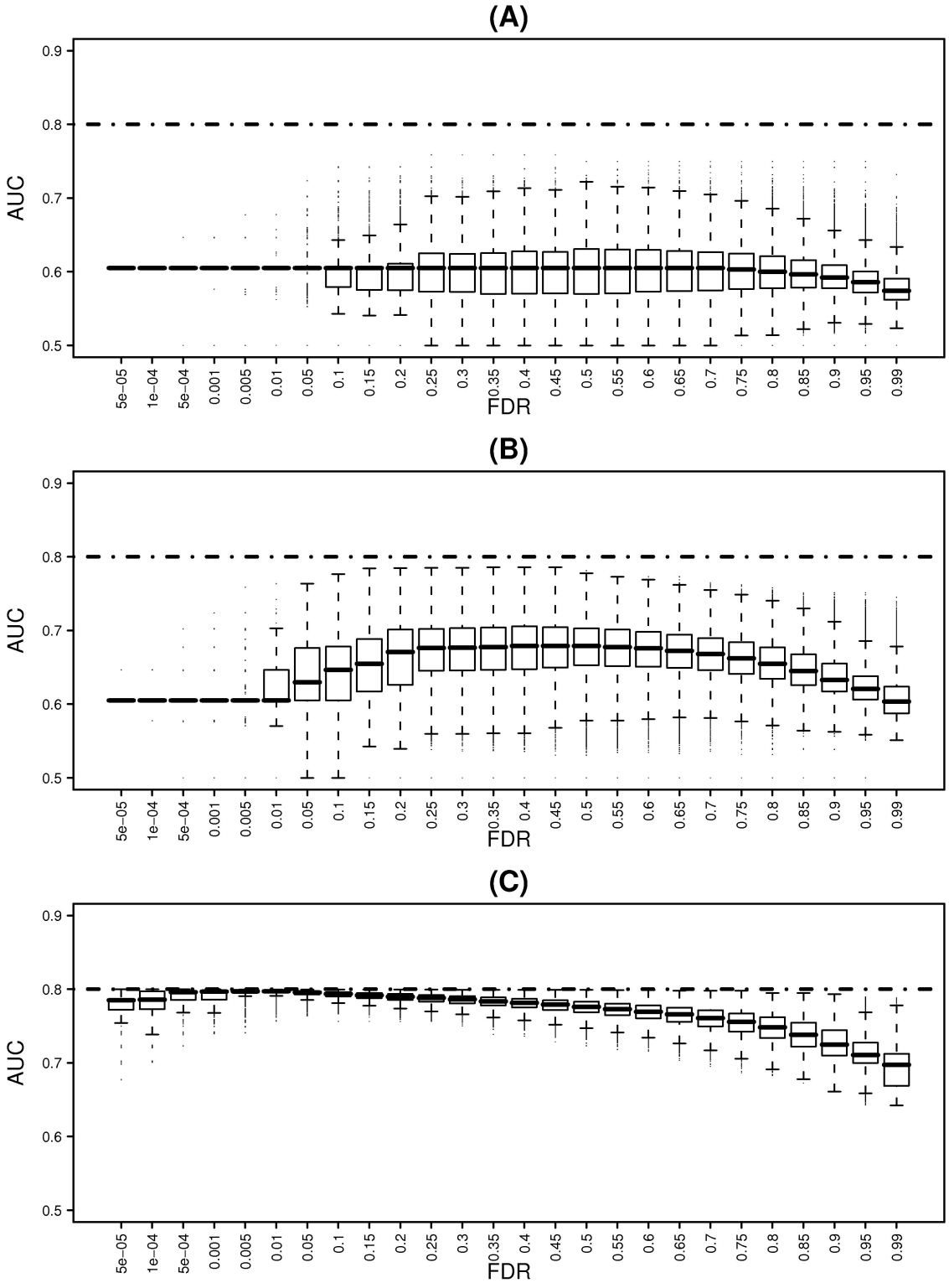

Figure 4.35: Boxplots of the area under the ROC-Curve (10000 simulated samples) for selection using the FDR approach assuming $m_e = 10$ alternatives markers among $m = 1000$ tested markers. The sample size per group was set to $n = 50$ (Figure (A)), 100 (Figure (B)) and 500 (Figure (C)). AUC_* is given as dotdashed horizontal line. Δ was set to 0.38.

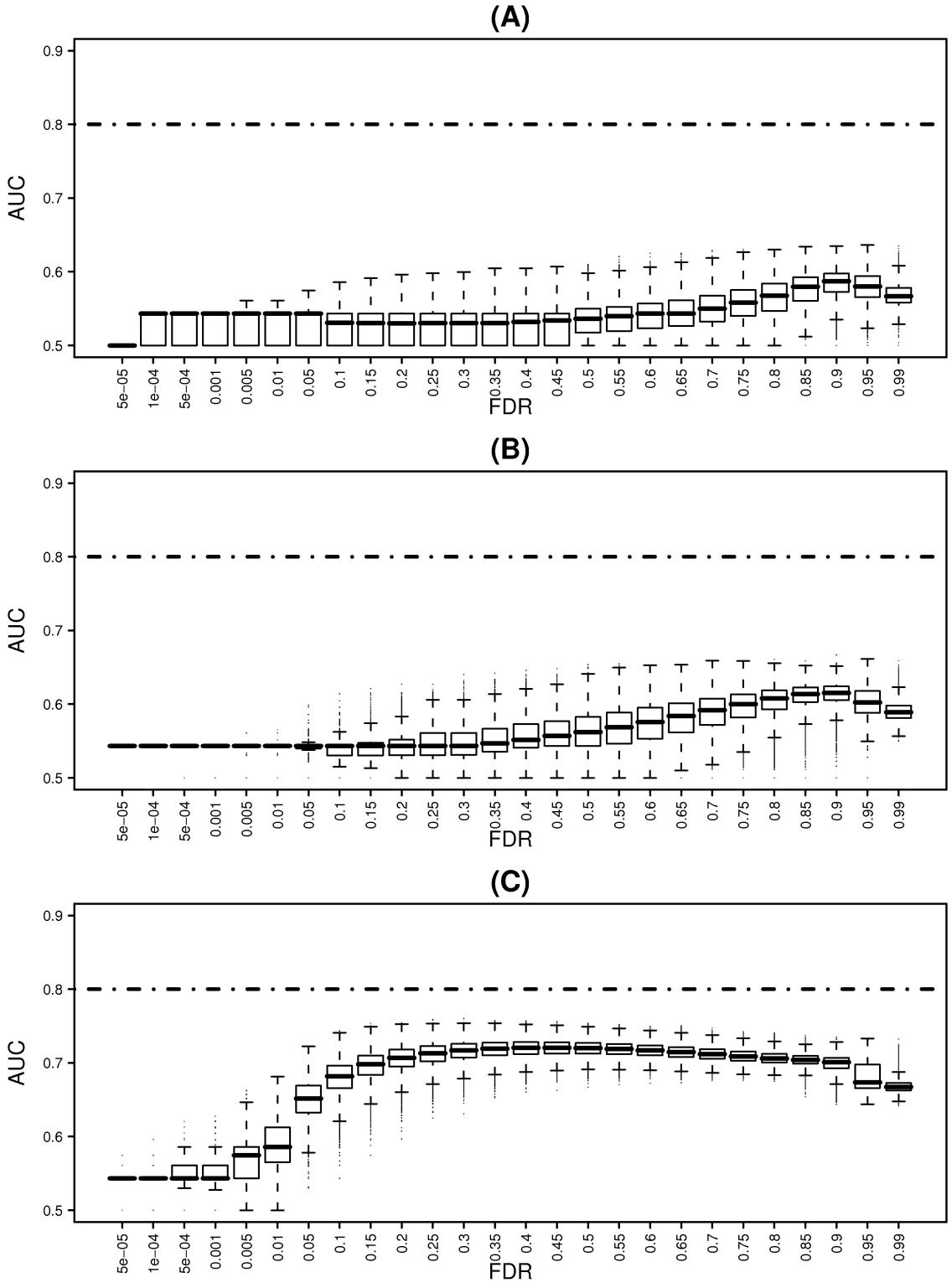

Figure 4.36: Boxplots of the area under the ROC-Curve (10000 simulated samples) for selection using the FDR approach assuming $m_e = 60$ alternatives markers among $m = 1000$ tested markers. The sample size per group was set to $n = 50$ (Figure (A)), 100 (Figure (B)) and 500 (Figure (C)). AUC_* is given as dotdashed horizontal line. Δ was set to 0.15.

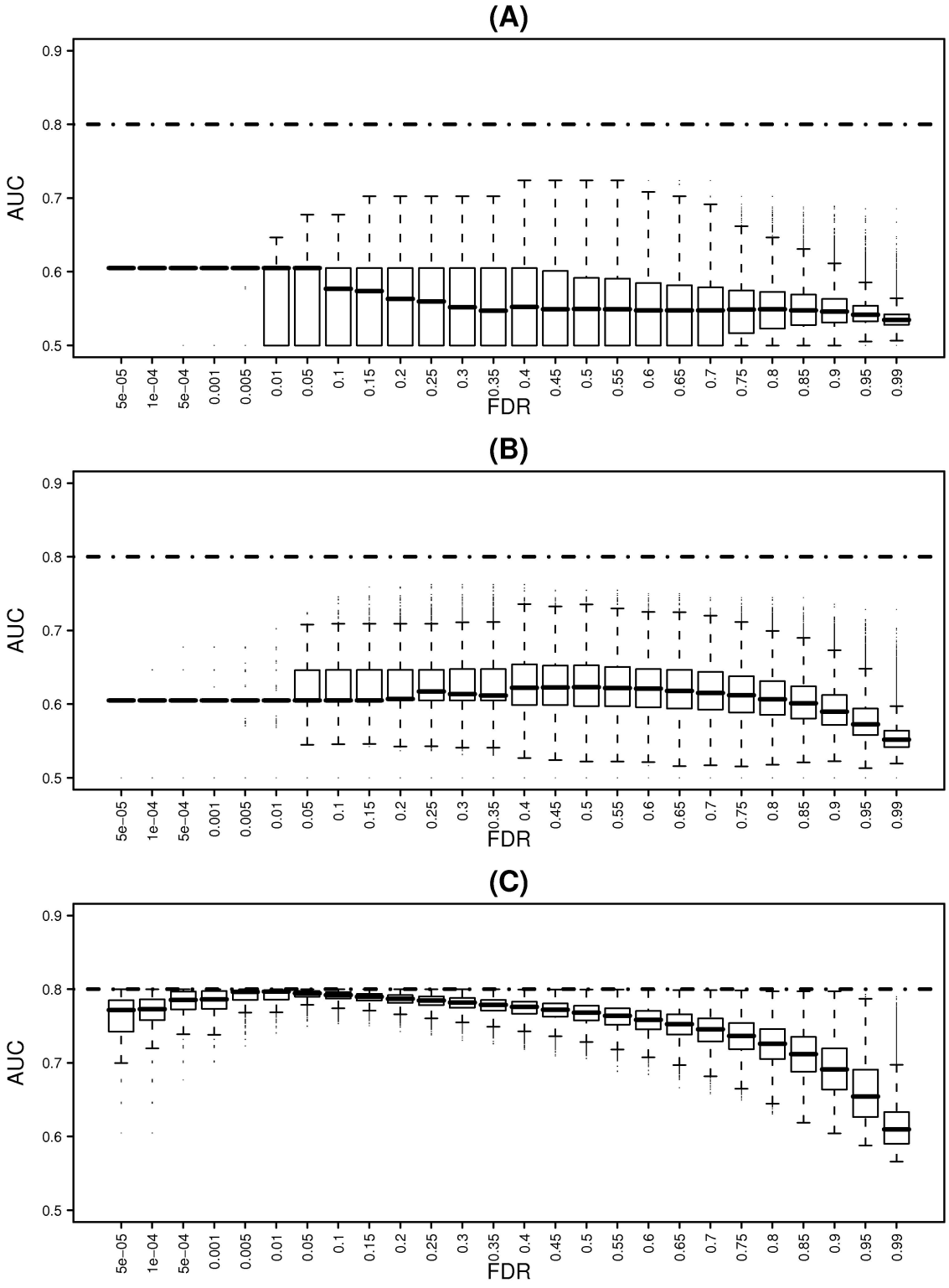

Figure 4.37: Boxplots of the area under the ROC-Curve (10000 simulated samples) for selection using the FDR approach assuming $m_e = 10$ alternatives markers among $m = 6000$ tested markers. The sample size per group was set to $n = 50$ (Figure (A)), 100 (Figure (B)) and 500 (Figure (C)). AUC_* is given as dotdashed horizontal line. Δ was set to 0.38.

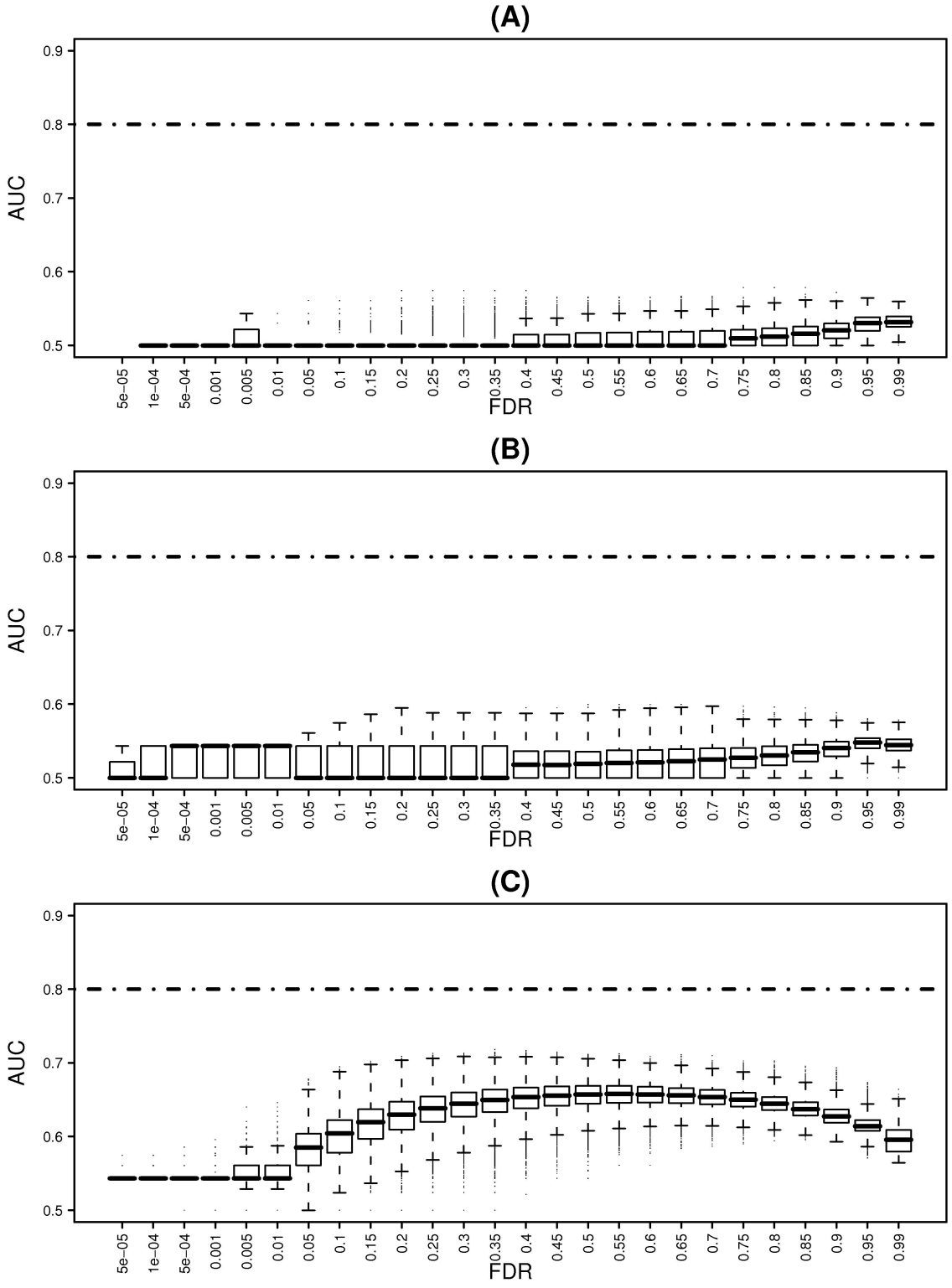

Figure 4.38: Boxplots of the area under the ROC-Curve (10000 simulated samples) for selection using the FDR approach assuming $m_e = 60$ alternatives markers among $m = 6000$ tested markers. The sample size per group was set to $n = 50$ (Figure (A)), 100 (Figure (B)) and 500 (Figure (C)). AUC_* is given as dotdashed horizontal line. Δ was set to 0.15.

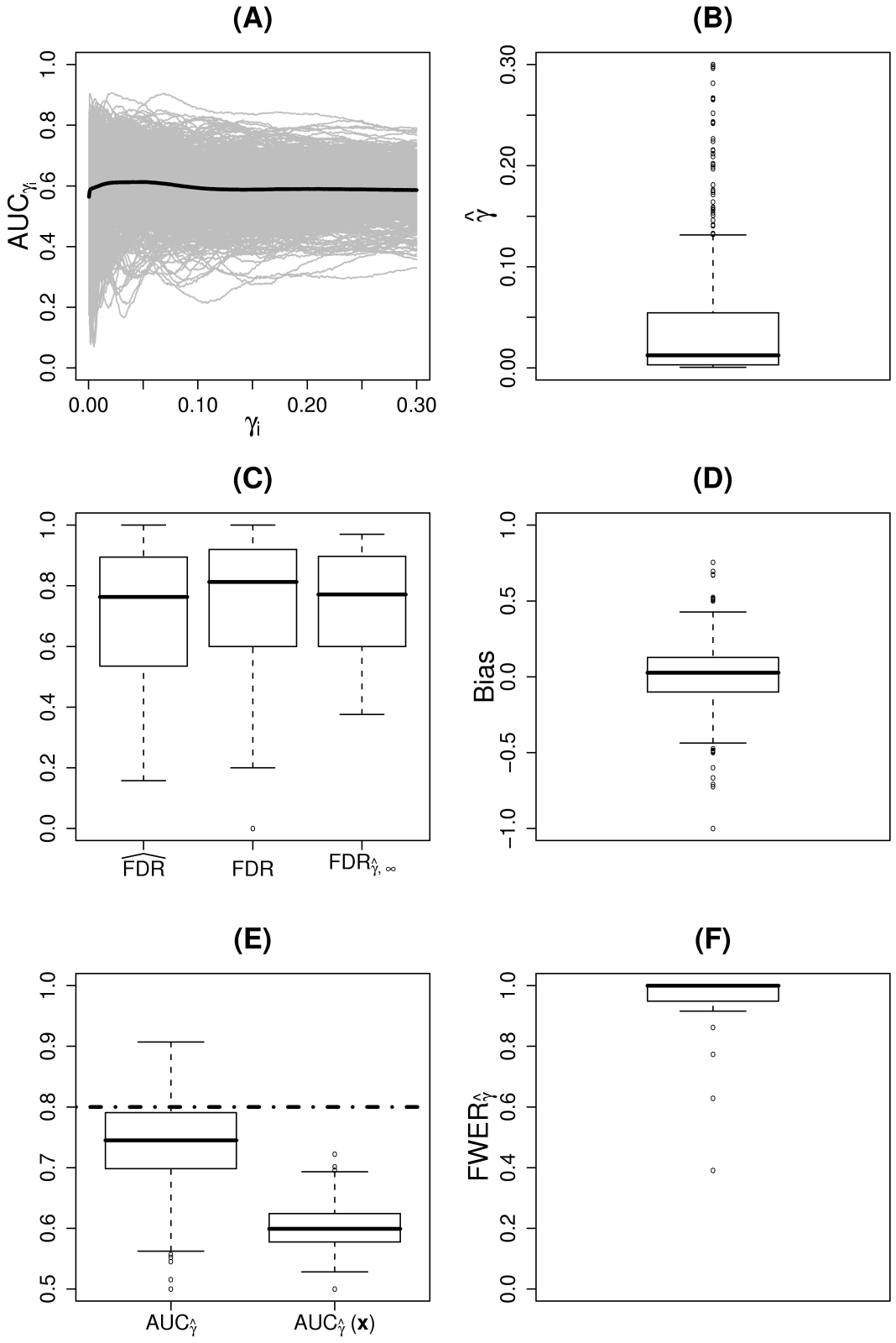

Figure 4.39: Simulation results (500 repetitions) of the Jackknife procedure applying the protected approach: AUC_{γ_i} as a function of γ_i (Figure (A)) as well as boxplots of the final selection boundary $\hat{\gamma}$ (B), the resulting estimated $\widehat{FDR}$, actual FDR and asymptotic $FDR_{\hat{\gamma}}$ (C), the bias of the estimated FDR (D), the jackknife $AUC_{\hat{\gamma}}$ and the future $AUC_{\hat{\gamma}}(\mathbf{x})$ (E) and $FWER_{\hat{\gamma}}$ (F) for the situation of $m_e = 10$, $m = 1000$, $n = 50$.

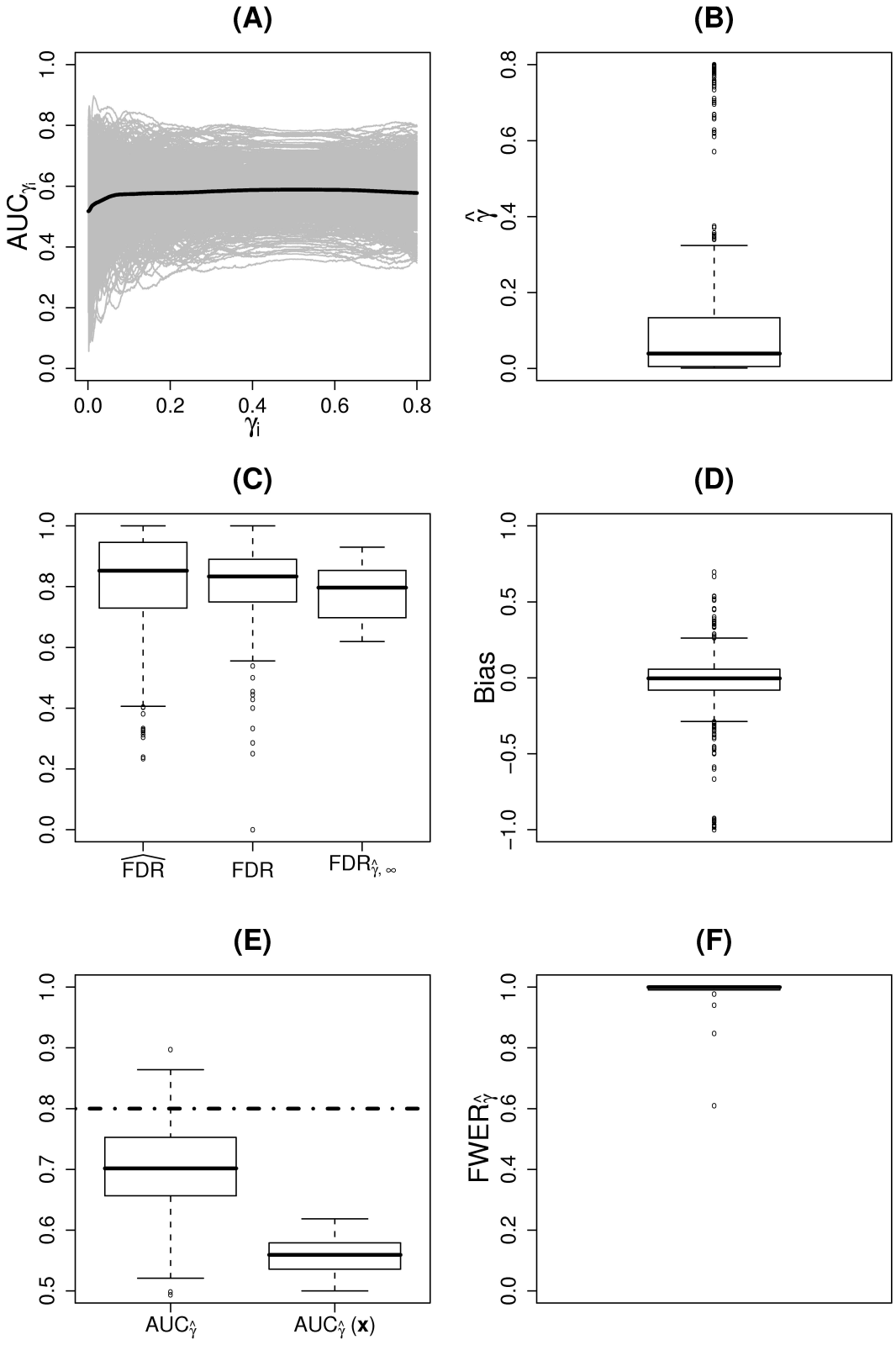

Figure 4.40: Simulation results (500 repetitions) of the Jackknife procedure applying the protected approach: AUC_{γ_i} as a function of γ_i (Figure (A)) as well as boxplots of the final selection boundary $\hat{\gamma}$ (B), the resulting estimated $\widehat{FDR}$, actual FDR and asymptotic $FDR_{\hat{\gamma}}$ (C), the bias of the estimated FDR (D), the jackknife $AUC_{\hat{\gamma}}$ and the future $AUC_{\hat{\gamma}}(\mathbf{x})$ (E) and $FWER_{\hat{\gamma}}$ (F) for the situation of $m_e = 60$, $m = 1000$, $n = 50$.

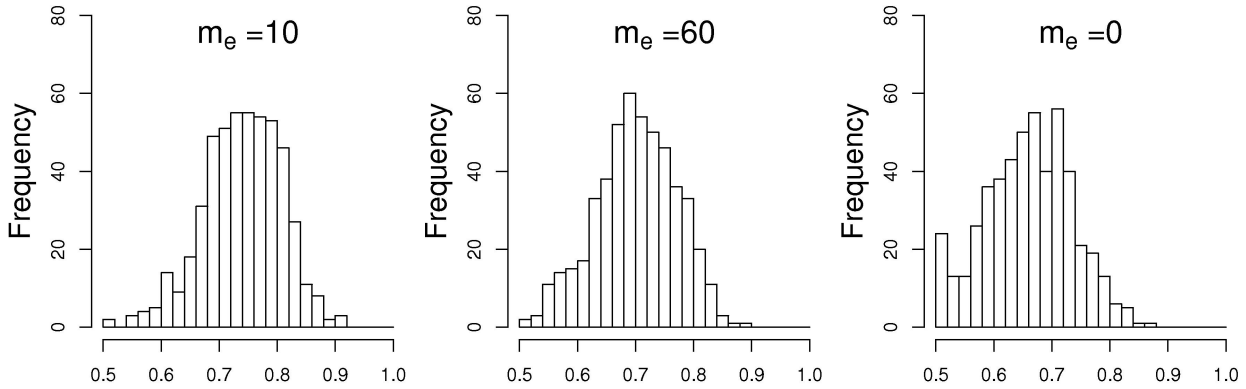

Figure 4.41: Distribution of the jackknife AUC for selection using FDR assuming a hypothetical AUC of 0.8 (500 simulation steps): $m_e = 10$, 60 or 0 among $m = 1000$ hypotheses. The sample size was set to $n = 50$ per group.

4.9 Discussion

If there is only a single marker candidate, to get a good prediction (ROC-Curve) the effect size has to be very large. Therefore only small samples are sufficient to identify this single marker by a statistical comparison between responders and non responders. But this is not the typical situation we face in such problems. Instead we generally are confronted with more than one candidate. Often there are very large numbers of candidates, few of them being effective with rather small effect sizes. Selection and estimation are often based on samples dramatically smaller than the number of candidates so that the asymptotic of model selection procedures does not apply. The estimates of the selected weights (and the ROC-curves) are biased and highly variable.

We performed simulations for different methods for selection of markers for future prediction. In the situation where the sample sizes are much smaller than the number of markers the simple method of additive scores following a selection of markers by multiple testing based on a FDR threshold general outperforms selection by forward stepwise logistic regression. The appearing problem of complete separation of data points results in selecting only a few alternatives for future prediction and thus to a poor performance in terms of AUC values for future independent patients. If the number of effective markers is rather small and they have sufficiently large effects, which, if all would be known, would lead to a large AUC, then

the selected scores may have good properties over a range of different FDR values used for selection. Under the alternative in general it seems to be preferable to use rather liberal selection criteria accepting that a certain number of nuisance markers is contained in a score to get the advantage of catching more effective ones. Even using the optimistic rule of simply selecting a pre-fixed number of "best" markers under the alternative may lead to good predictions in terms of the AUC of the resulting score. For large samples the predictive ability of the estimated score does not depend strongly on the number of selected markers. For large sample sizes the weights for ineffective markers contained in the score are estimated more precisely and, despite of the selection procedure, will tend to be the ones close to zero with a small contribution overall. Hence the performance of the score will not vary much for different numbers of ineffective markers contained in the score: More liberal selection criteria will lead to scores containing more nuisance markers (with low weights) but also more effective ones (with large weights). This mirrors the fact that asymptotically for large sample sizes, multiple test based selection procedures may be consistent procedures for model selection. It also has to be mentioned, that if a very large number of effective markers is expected working together with rather small individual effect sizes and only small sample sizes are available, the two selection methods based on univariate tests also perform worse, although leading to a larger AUC for future prediction than using the forward logistic regression.

The cruical scenario in the small sample case is the global null hypothesis: There are no effective markers at all and hence any selection will lead to completely uninformative prediction scores. To protect against erroneous selection in this situation, e.g. the FDR applied for selection should be rather small. Under the global null hypothesis control of the FDR also controls the probability of the selection of any markers. Under the alternative, however, we found that rather larger FDR values should be used for selection. The unprotected approach will always select nuisance scores from the data.

One way to determine the FDR-value to be applied for selection in a concrete sample (or simply the number k of the best markers to be selected) in order to achieve good prediction by a prognostic score in terms of the AUC is to estimate the selection boundaries by jack-

knife methodology. For a grid of different selection thresholds we repeated the estimation of the score in the n^2 samples resulting from leaving out all n^2 possible pairs of a single non-responder and a single responder respectively. We then estimated the jackknife based AUC_{γ_i} by counting how often in the n^2 pairs left out from the respective training samples the score derived from the corresponding training sample has the right order in magnitude between the non-responder and the responder. The selection thresholds leading to the largest jackknife based AUC_{γ_i} was used for selection in the total sample. This method seems to work if we really are in a situation that we deal with markers with a high prognostic potential which leads to rather large jackknife estimates of the AUC, whereas under the global null hypothesis these estimates concentrate around 0.5.

However the situation gets much worse if we look at scenarios when the markers are not sufficiently effective that the optimal score (if known) would lead to a ROC-curve crossing the point with sensitivity and specificity of 0.9 (with a theoretically best achievable $AUC_* = 0.965$ which was the benchmark in most of our investigations). There may be biological situations where markers of that predictive potential can be expected. For AUC_* close to values of 0.8, which are values, e.g., achievable in predicting hospital mortality from scores based on a set of variable measured in patients at admission to an intensive care unit, the selection procedure will lead to poor scores. Also jackknife estimation procedure of the AUC to determine the selection thresholds will require much larger samples in order to produce good selection thresholds.

With this cross validation method we may achieve several goals:

1. We determine an optimal selection threshold for selection of markers to be used in a prediction score for future sample units which provides a good performance in terms of the ROC-curve.

2. We get a positively biased estimate of the AUC which is closer to the true AUC for prediction the larger the effect and sample sizes.

3. If the estimate of the AUC is small this may be an indication that in a specific sample

we are close to the global null hypothesis or the effect sizes are to small for the given sample size.

4. We also get an estimate of the FDR among the selected markers which is close to the true FDR (with a direction depending on the magnitude of the FDR).

Our findings show that simple methods can lead to well performing prediction scores even in rather small samples, given that we deal with a problem where markers with noticeable effects are involved. However, they also tell us that there is no such thing as a free lunch in a statistically odd problem of dealing with large numbers of variables considerably exceeding sample sizes.

In this thesis we did not consider the situation of unknown variances or correlation between the hypotheses since it would go beyond the scope of this work. However these topics are object of our future scientific work.

Acknowledgement

I am grateful to my supervisor Peter Bauer for his support, his assistance and his patience during the years.
I also want to thank Sonja Zehetmayer and Martin Posch for discussing many aspects of this work and for many helpful comments as well as Franz König for his advice regarding SAS codes. Last but not least I would like to express my gratitude to my parents Johann and Maria Goll as well as to Franklin, Ferdinand and Andreas Graf for their support and their continuous encouragement to remain defiant.

This work was supported by the Austrian FWF-Fund no. P18698-n15.

A Abstract

Motivation:

Two problems that arise in the genomic or proteomic context are to find differentially expressed genes (proteins) among a large number of hypotheses and to find prognostic scores to predict a clinical outcome for future patients. Referring to the first problem, it has been shown, that two-stage pilot and integrated designs are powerful tools for investigating large numbers of hypotheses. In modern genetic studies often the costs per observation differ between stages, because specific experimental devices have to be produced at higher costs per measurement for the selected markers in the second stage. There is also an increasing focus on using a less accurate assay in early stages and more accurate, but more expensive ones in later stages for cost reasons. Asymptotically optimal two-stage designs controlling the Family Wise Error or False Discovery Rate are considered when costs and effect sizes per measurement differ between stages and total costs are constrained.

Investigating the second mentioned problem it seems that for a lot of medical research reported in this field there has been insufficient awareness of the statistical properties of the resulting prognostic scores. We looked at the statistical properties of such scores in terms of how well they can predict the outcome of a medical therapy in future patients in terms of the area under the receiver operating characteristics (AUC).

Results:

For the practically relevant case that the same method is applied at both stages but designing the second stage measurements raises extra costs, two-stage designs are more powerful than the single-stage design even for large costs ratios. The power of the optimal pilot

and integrated two-stage designs generally are similar, however the integrated approach is less sensitive even to severe design misspecifications in the planning phase. Depending on the cost and effect size ratios between the measurements it is generally more powerful to apply two-stage procedures using one measurement method at both stages. Switching from a low-cost standard method to a high-cost improved method may only be advisable if there is lack of resources, so that the first-stage sample size for the high-cost method would be too small.

Selection and estimation are often based on samples dramatically smaller than the number of candidates so that the asymptotic of model selection procedures (sample size goes to infinity) does not apply. The estimates of the selected weights are biased and highly variable. If there are in fact effective markers among many candidates, prediction seems to be better if the selection procedures allows to include some ineffective markers in the score. If there are no effective markers at all, that means the global null hypothesis is true, then applying a liberal selection procedure tends to create pseudo scores which have no prognostic value at all. The best threshold to be used in selection of the prediction score which provides the largest AUC varies over different parameter constellations (varying number of tested markers, proportion of alternatives or sample size). We considered cross validation to determine the optimal selection criterion in a specific sample. For that purpose we used a modified jackknife procedure. This procedure allows choosing an appropriate selection criterion for constructing a prediction score and at the same time provides an estimate for the extent of false positive decisions. Moreover, this procedure by leading to low jackknife AUCs may indicate that in a particular data set we are close to the global null hypotheses or the effects are to small for the given sample size.

B Kurzfassung

Motivation:

In genetischen Studien trifft man häufig auf zwei Probleme: Erstens versucht man Gene (oder Proteine) zu finden die zum Beispiel bei bestimmten Krankheiten im Vergleich zu gesunden Personen verschieden ausgeprägt sind. Andererseits versucht man aus solchen Genen (Proteinen) prognostische Scores für den klinischen Ausgang zum Beispiel einer Therapie eines Patienten zu finden. Schon in früheren Publikationen wurde gezeigt, dass Zwei-Stufen Pläne, wie das Pilot oder das Integrated Design geeignete Methoden sind um eine große Anzahl von Hypothesen zu testen, wie es in solchen Studien der Fall ist. In genetischen Studien kann es vorkommen, dass die Kosten pro Beobachtung zwischen den beiden Stufen unterschiedlich sind, wenn zum Beispiel spezielle Chips für die Untersuchung von selektierten Genen angefertigt werden müssen. In neueren Studien kommt es auch immer öfter vor, dass in der ersten Stufe ein billiges Standardverfahren und in der zweiten Stufe teureres, aber dafür genaueres Verfahren verwendet wird. Für solche Zwei-Stufen Pläne, in denen Kosten und Methoden zwischen den Stufen variieren, werden asymptotisch optimale Parameter untersucht. Es soll entweder die False Discovery Rate oder der Family Wise Error einhalten werden.

Bezugnehmend auf das zweite angeführte Problem gibt es für viele medizinische Untersuchungen in diesem Gebiet nur mangelhafte Erkenntnisse über die statistischen Eigenschaften prognostischer Scores, die auf Basis von genetischen Datensätzen entwickelt werden. Solche statistische Eigenschaften, das heißt wie gut ein Score, der aufgrund eines Trainigsdatensatzes erstellt wurde, den Ausgang einer Therapie eines zukünftigen Patienten vorhersagt, werden untersucht. Es wird dabei die "area under the receiver operating characteristic" (AUC) als Kriterium verwendet.

Ergebnisse:

Unterscheiden sich nur die Kosten pro Beobachtung zwischen den Stufen, zahlt sich auch bei einem beachtlichen Kostenverhältnis zwischen den beiden Stufen die Anwendung von Zwei-Stufen Verfahren aus (im Sinne von höherer Power). Die Power der optimalen Pilot und Integrated Designs ist zwar ähnlich hoch, das Integrated Design ist allerdings robuster gegen Fehleinschätzungen der Parameter in der Planungsphase. Im Allgemeinen führen Zwei-Stufen Pläne, die die gleiche Methode in beiden Stufen verwenden, zu einer höheren Power (in Abhängigkeit von Kosten- und Effektgrößenverhältis), als Pläne, bei denen die Methode in der zweiten Stufe geändert wird. Letztere Pläne zahlen sich nur aus, wenn die teurere Methode aufgrund von fehlenden finanziellen Mitteln in der ersten Stufe nicht verwendet werden kann.

Die Selektion von wirksamen Genen (Proteinen) und somit die Aufstellung von prognostischen Scores hängt sehr oft von den Trainigsdaten ab, bei denen die Stichprobengröße oft drastisch kleiner ist als die Anzahl der untersuchten Hypothesen, so dass die Asymptotik (Stichprobenumfang geht gegen unendlich) von Modellselektionsprozeduren nicht mehr stimmt. Diese Scores sind daher verzerrt und sehr variabel. Wenn tatsächlich mehrere effektive Marker existieren stellt es sich heraus, dass es besser ist liberalere Selektionskiterien zu verwenden und somit auch einige (oft sogar viele) ineffektive Marker in den Score mit aufzunehmen. Gilt allerdings die Globale Nullhypothese führen solche liberalen Kriterien dazu, dass Pseudo-Scores erstellt werden, die keinen prognostischen Wert haben. Das beste Selektionskriterium um einen prognostischen Score mit einer hohen AUC zu erzeugen ist allerdings von Parametern wie der Anzahl der insgesamt gestesteten Hypothesen, der Stichprobengröße oder dem Anteil der effektiven Marker abhängig. Wir testeten eine modifizierte Jackknife Methode um das optimal Selektionskriterium zu finden. Es stellt sich heraus, dass diese Methode ein geeignetes Selektionskriterium findet und zusätzlich Informationen über die Anzahl der falsch-positiven Entscheidungen liefert. Darüber hinaus bekommt man durch kleine Jackknife-AUC Werte einen Hinweis darauf, dass man sich in einem spezifischen Datensatz in der "Nähe" der globale Nullhypothese befindet oder dass die Effekte für den gegebenen Stichprobenumfang zu kein sind.

C Curriculum Vitae

Personal data:

- Name: Alexandra Goll
- Date of Birth: 20.08.1979
- Place of Birth: Vienna
- Nationality: Austria
- Parents: Ing. Johann and Maria Goll

Education:

- 1993-1998 Commercial academy, Korneuburg, Niederösterreich
  Graduation: June, 1998
- 1998-2003 Academic studies in Statistics, Graduation: October, 2003
- since WT 2000: Academic studies in Mathematics, completion first part: March, 2004
- since ST 2005: Doctoral studies in Statistics

Work experience:

- since October, 2003: Part-time job at the Institute of Medical Statistics, Core Unit for Medical Statistics and Informatics, Medical University of Vienna
- since April, 2005: additional part-time job as doctoral student at the Institute of Medical Statistics: current FWF-Project Nr.: P18698-n15
- ST 2005: Lecturer at the Medical University Vienna: "SSM2: Methoden der Medizininischen Wissenschaften"
- ST 2006: Lecturer at the Medical University Vienna: "SSM3: Methodenseminar Statistik"
- ST 2006: External lecturer at the University of Vienna: "Übungen zu Mathematik 2 für Statistiker und Volkswirtschaftler"

Talks about the thesis:

- International Conference on Multiple Comparison Procedures (MCP), Juli 2007 in Vienna, Austria
- ROeS Seminar, September 2007 in Bern, Switzerland
- Statistics and Life Sciences: Perspectives and Challenges (LIFESTAT), March 2008 in Munich, Germany

Papers concerning the thesis and clinical cooperations:

- Goll, A. and Bauer, P. (2007). Two-stage designs applying methods differing in costs, *Bioinformatics*, 23: 1519-1526.
- Zehetmayer,S., Goll, A., Bauer,P. and Posch,M. (2007). Step by Step: mehr Effizienz mit neuen Studiendesigns, *Biospektrum*, 7: 754-755.
- Sabeti-Aschraf, M., Serek, M., Pachtner, T., Auner, K., Machinek, M., Geisler, M. and Goll A. (2008). The Enduro motorcyclist's wrist and other overuse injuries in competitive Enduro motorcyclists: a prospective study. *Scand J Med Sci Sports*, to appear
- Krebs, I., Ansari-Shahrezaei, S., Goll, A. and Binder, S. (2008). Activity of neovascular lesions treated with bevacizumab: comparison between optical coherence tomography and fluorescein angiography. *Graefes Arch Clin Exp Ophthalmol*, to appear
- Rabenlehner, D., Stanzel, B.V., Krebs, I., Binder, S. and Goll, A. (2008). Reduction of iatrogenic RPE lesions in AMD patients: evidence for wound healing? *Graefes Arch Clin Exp Ophthalmol*, 246: 345-352.
- Krebs, I., Krepler, K., Stolba, U., Goll, A. and Binder, S. (2008). Retinal angiomatous proliferation: combined therapy of intravitreal triamcinolone acetonide and PDT versus PDT alone. *Graefes Arch Clin Exp Ophthalmol*, 246: 237-243.
- Winkler, W., Zellner, M., Diestinger, M., Babeluk, R., Marchetti, M., Goll, A., Zehetmayer, S., Bauer, P., Rappold, E., Miller, I., Roth, E., Allmaier, G. and Oehler, R. (2008). Biological variation of the platelet proteome in the elderly population and its implication for biomarker research. *Mol Cell Proteomics*, 7: 193-203.
- Sabeti, M., Dorotka, R., Goll, A., Gruber, M. and Schatz, K.D. (2007). A comparison of two different treatments with navigated extracorporeal shock-wave therapy for calcifying tendinitis - a randomized controlled trial. *Wien Klin Wochenschr*, 119: 124-128.
- Dorotka, R., Sabeti, M., Jimenez-Boj, E., Goll, A., Schubert, S. and Trieb, K. (2006). Location modalities for focused extracorporeal shock wave application in the treatment of chronic plantar fasciitis. *Foot Ankle Int*, 27: 943-947.
- Stacher, G., Lenglinger, J., Eisler, M., Hoffmann, M., Goll, A., Bergmann, H. and Stacher-Janotta, G. (2006). Esophageal acid exposure in upright and recumbent postures: roles of lower esophageal sphincter, esophageal contractile and transport function, hiatal hernia, age, sex, and body mass. *Dig Dis Sci*, 51: 1896-1903.
- Zehetgruber, H., Grübl, A., Goll, A., Schwameis, E., Wurnig, C. and Giurea, A. (2005). Prevention of heterotopic ossification after THA with indomethacin: analysis of risk factors. *Z Orthop Ihre Grenzgeb*, 143: 631-637.
- Krebs, I., Binder, S., Stolba, U., Glittenberg, C., Brannath, W. and Goll, A. (2005) Choroidal neovascularization in pathologic myopia: three-year results after photodynamic therapy. *Am J Ophthalmol*, 140: 416-425.
- Krebs, I., Binder, S., Stolba, U., Schmid, K., Glittenberg, C., Brannath, W. and Goll, A. (2005). Optical coherence tomography guided retreatment of photodynamic therapy. *Br J Ophthalmol*, 89: 1184-1187.
- Sabeti-Aschraf, M., Dorotka, R., Goll, A. and Trieb, K. (2005). Extracorporeal shock wave therapy in the treatment of calcific tendinitis of the rotator cuff. *American Journal of Sports Medicine*, 33: 1365-1368.
- Windberger, U., Grohmann, K., Goll, A., Plasenzotti, R. and Losert, U (2005). Fetal and juvenile animal hemorheology. *Clin Hemorheol Microcirc*. 32: 191-197.

D R-Code

The R-Programs for the two-stage pilot and integrate designs controlling the FWER or FDR include the following parameters.

Parameters:

TotalC	... fixed overall total costs C
c2	... cost ratio c_2 between both stages
r	... fraction r of C used in the first stage
m1	... number of hypotheses m_1
pi0	... proportion of true null hypotheses π_0 among all m_1 hypotheses
alpha	... targeted false discovery rate α
ga1	... first stage selection boundary γ_1
ga2	... second stage rejection boundary γ_2 for the pilot design
ga	... second stage rejection boundary γ for the integrated design
delta	... effect size Δ
kr	... effect size ratio k between both stages
n1	... first stage sample size n_1
n2	... second stage sample size n_2
dparms	... =c(r,ga1) vector containing the two design parameters r and γ_1
ivga1	... initial value of γ_1 for optimization
ivr	... initial value of r for optimization

D.1 The pilot design controlling the FWER

This in an R-Program to compute the asymptotically optimal design parameters r and γ_1 as well as the power $\prod_p$ for a TWO-STAGE PILOT DESIGN controlling the FWER.

Functions:

optpilotfwe	... function to compute (-1)*(asymptotically optimal power) -> minimization criteria for optimization
pilotfwe	... function to compute the asymptotically optimal design parameters r and γ_1

R-Code:

```
optpilotfwe<-function(dparms,m1,pi0,delta,kr,alpha,TotalC,c2)
{
r<-dparms[1]
ga1<-dparms[2]
n1<-r*TotalC/m1
powerst1<-1-pnorm(qnorm(1-ga1)-delta*sqrt(n1))
m2<-m1*(1-pi0)*powerst1+m1*pi0*ga1
ga2<-alpha/m2
n2<-(1-r)*TotalC/c2/m2
powerst2<-1-pnorm(qnorm(1-ga2)-delta*kr*sqrt(n2))
-powerst1*powerst2
}

pilotfwe<-function(m1,pi0,delta,kr,alpha,TotalC,c2,ivr,ivga1)
{

optpilot <-optim(c(ivr,ivga1),fn=optpilotfwe,m1=m1,pi0=pi0,delta=delta,kr=kr,
        alpha=alpha,TotalC=TotalC,c2=c2,method="L-BFGS-B",
        lower=c(0.00000001,0.00000001),upper=c(0.999999,0.999999))

r<-optpilot[[1]][1]
ga1<-optpilot[[1]][2]
totalpower<–optpilot[[2]]
result<-round(c(r,ga1,totalpower),3)
names(result)<-c("r","gamma1","Power")
result
}
```

D.2 The pilot design controlling the FDR

This in an R-Program to compute the asymptotically optimal design parameters r and γ_1 as well as the power $\prod_p$ for a TWO-STAGE PILOT DESIGN controlling the FDR.

Functions:

pfdr.fun	... function to compute the asymptotic FDR alpha as a function of γ_1 and γ_2
bspilotfdr	... function to compute γ_2 for given γ_1 and α
optpilotfdr	... function to compute (-1)*(asymptotically optimal power) ->minimization criteria for optimization
pilotfdr	... function to compute the asymptotically optimal design parameters r and γ_1

R-Code:

```
pfdr.fun<-function(pi0,ga1,ga2,n1,n2,delta,kr)
{
powerst1<-1-pnorm(qnorm(1-ga1),mean=delta*sqrt(n1),sd=1)
powerst2<-1-pnorm(qnorm(1-ga2),mean=delta*kr*sqrt(n2),sd=1)
alpha<-pi0*ga1*ga2/(pi0*ga1*ga2+(1-pi0)*powerst1*powerst2)
alpha
}

bspilotfdr<-function(alpha,pi0,ga1,n1,n2,delta,kr)
{
ga2a<-0.0000000001
ga2b<-0.999999
i<-100
while (i>0) {

  ga2c<-(ga2a+ga2b)/2
  alphaze<-pfdr.fun(pi0=pi0,ga1=ga1,ga2=ga2c,n1=n1,n2=n2,delta=delta,kr=kr)
  ga2a<-ifelse(alphaze<=alpha,ga2c,ga2a)
  ga2b<-ifelse(alphaze<=alpha,ga2b,ga2c)
  i<-i-1 }

ga2c
}
```

```
optpilotfdr<-function(dparms,m1,pi0,delta,kr,alpha,TotalC,c2)
{
r<-dparms[1]
ga1<-dparms[2]
n1<-r*TotalC/m1
powerst1<-1-pnorm(qnorm(1-ga1),mean=delta*sqrt(n1),sd=1)
m2<-m1*(1-pi0)*powerst1+m1*pi0*ga1
n2<-(1-r)*TotalC/c2/m2
ga2<-bspilotfdr(alpha,pi0,ga1,n1,n2,delta,kr)
powerst2<-1-pnorm(qnorm(1-ga2),mean=delta*kr*sqrt(n2),sd=1)
-powerst1*powerst2
}

pilotfdr<-function(m1,pi0,delta,kr,alpha,TotalC,c2,ivr,ivga1)
{

optpilot<-optim(c(ivr,ivga1),fn=optpilotfdr,m1=m1,pi0=pi0,delta=delta,kr=kr,
        alpha=alpha,TotalC=TotalC,c2=c2,method="L-BFGS-B",
        lower=c(0.00000001,0.00000001),upper=c(0.999999,0.999999))

r<-optpilot[[1]][1]
ga1<-optpilot[[1]][2]
totalpower<–optpilot[[2]]
result<-round(c(r,ga1,totalpower),3)
names(result)<-c("r","gamma1","Power")
result
}
```

D.3 The integrated design controlling the FWER

This in an R-Program to compute the asymptotically optimal design parameters r and γ_1 as well as the power $\prod_{int}$ for a TWO-STAGE INTEGRATED DESIGN controlling the FWER.

Functions:

powerintfwe	... function to compute the asymptotic power of the integrated design (without integration over z)
gintfwe	... function to compute γ_s for given γ and γ_1 of the integrated design (without integration over z)
bsintfwe	... function to compute γ for given γ_1 and α
optintfwe	... function to compute (-1)*(asymptotically optimal power) -> minimization criteria for optimization
integratedfwe	... function to compute the asymptotically optimal design parameters r and γ_1

R-Code:

```
powerintfwe<-function(z,ga,n1,n2,delta,kr)
{
w1<-n1/(n1+n2*kr^2)

(1-pnorm ((qnorm(1-ga)-sqrt(w1)*z)/sqrt(1-w1),mean=(sqrt(n2)*delta),sd=1))
        *dnorm(z,mean=sqrt(n1)*delta*kr,sd=1)

}

gintfwe<-function(z,ga,n1,n2,kr)
{
w1<-n1/(n1+n2*kr^2)
(1-pnorm((qnorm(1-ga)-sqrt(w1)*z)/sqrt(1-w1)))*dnorm(z)
}

bsintfwe<-function(alpha,m1,ga1,n1,n2,delta,kr)
{
gas<-alpha/m1
gaa<-0.0000000001
gab<-0.999999
i<-100
while (i>0) {

  gac<-(gaa+gab)/2
  gammaze<-integrate(gintfwe,lower=qnorm(1-ga1),upper=Inf,ga=gac,n1=n1,n2=n2,kr=kr)[[1]]
  gaa<-ifelse(gammaze<=gas,gac,gaa)
  gab<-ifelse(gammaze<=gas,gab,gac)
  i<-i-1 }

gac
}
```

```
optintfwe<-function(dparms,delta,kr,alpha,m1,TotalC,c2,pi0)
{
r<-dparms[1]
ga1<-dparms[2]
n1<-TotalC*r/m1
m2<-m1*(pi0*ga1+(1-pi0)*(1-pnorm(qnorm(1-ga1),mean=delta*sqrt(n1))))
n2<-(1-r)*TotalC/c2/m2
ga<-bsintfwe(alpha=alpha,m1=m1,ga1=ga1,n1=n1,n2=n2,delta=delta,kr=kr)
powerint<-integrate(powerintfwe,qnorm(1-ga1),Inf,ga=ga,n1=n1,n2=n2,delta=delta,kr=kr)[[1]]
-powerint
}

integratedfwe<-function(m1,pi0,delta,kr,alpha,TotalC,c2,ivr,ivga1)
{

optint<-optim(c(ivr,ivga1),fn=optintfwe,m1=m1,pi0=pi0,delta=delta,kr=kr,
      alpha=alpha,TotalC=TotalC,c2=c2,method="L-BFGS-B",
      lower=c(0.00000001,0.00000001),upper=c(0.999999,0.999999))

r<-optint[[1]][1]
ga1<-optint[[1]][2]
totalpower<- -optint[[2]]
result<-round(c(r,ga1,totalpower),3)
names(result)<-c("r","gamma1","Power")
result
}
```

D.4 The integrated design controlling the FDR

This in an R-Program to compute the asymptotically optimal design parameters r and γ_1 as well as the power $\prod_{int}$ for a TWO-STAGE INTEGRATED DESIGN controlling the FDR.

Functions:

powerintfdr	... function to compute the asymptotic power of the integrated design (without integration over z)
gintfdr	... function to compute γ_s for given γ and γ_1 of the integrated design (without integration over z)
intfdr.fun	... function to compute the asymptotic FDR alpha as a function of γ and γ_1
bsintfdr	... function to compute γ for given γ_1 and α
optintfdr	... function to compute (-1)*(asymptotically optimal power) -> minimization criteria for optimization
integratedfdr	... function to compute the asymptotically optimal design parameters r and γ_1

R-Code:

```
powerintfdr<-function(z,ga,n1,n2,delta,kr)
{
w1<-n1/(n1+n2*kr^2)

(1-pnorm((qnorm(1-ga)-sqrt(w1)*z)/sqrt(1-w1),mean=(sqrt(n2)*kr*delta),sd=1))
        *dnorm(z,mean=sqrt(n1)*delta,sd=1)

}

gintfdr<-function(z,ga,n1,n2,kr)
{
w1<-n1/(n1+n2*kr^2)
(1-pnorm((qnorm(1-ga)-sqrt(w1)*z)/sqrt(1-w1)))*dnorm(z)
}

intfdr.fun<-function(pi0,ga1,ga,n1,n2,delta,kr)
{
gs<-integrate(gintfdr,lower=qnorm(1-ga1),upper=Inf,ga=ga,n1=n1,n2=n2,kr=kr)[[1]]

powint<-integrate(powerintfdr,lower=qnorm(1-ga1),upper=Inf,ga=ga,n1=n1,n2=n2,
      delta=delta,kr=kr)[[1]]

alpha<-pi0*gs/(pi0*gs+(1-pi0)*powint)
alpha
}
```

```
bsintfdr<-function(alpha,pi0,ga1,n1,n2,delta,kr)
{
gaa<-0.0000000001
gab<-0.999999
i<-100
while (i>0) {

   gac<-(gaa+gab)/2
   alphaze<-intfdr.fun(pi0=pi0,ga1=ga1,ga=gac,n1=n1,n2=n2,delta=delta,kr=kr)
   gaa<-ifelse(alphaze<=alpha,gac,gaa)
   gab<-ifelse(alphaze<=alpha,gab,gac)
   i<-i-1 }

gac
}

optintfdr<-function(dparms,delta,kr,alpha,m1,TotalC,c2,pi0)
{
r<-dparms[1]
ga1<-dparms[2]
n1<-TotalC*r/m1
m2<-m1*(pi0*ga1+(1-pi0)*(1-pnorm(qnorm(1-ga1),mean=delta*sqrt(n1))))
n2<-(1-r)*TotalC/c2/m2
ga<-bsintfdr(alpha=alpha,pi0=pi0,ga1=ga1,n1=n1,n2=n2,delta=delta,kr=kr)
powerint<-integrate(powerintfdr,qnorm(1-ga1),Inf,ga=ga,n1=n1,n2=n2,delta=delta,kr=kr)[[1]]
-powerint
}

integratedfdr<-function(m1,pi0,delta,kr,alpha,TotalC,c2,ivr,ivga1)
{

optint<-optim(c(ivr,ivga1),fn=optintfdr,m1=m1,pi0=pi0,delta=delta,kr=kr,
      alpha=alpha,TotalC=TotalC,c2=c2,method="L-BFGS-B",
      lower=c(0.00000001,0.00000001),upper=c(0.999999,0.999999))

r<-optint[[1]][1]
ga1<-optint[[1]][2]
totalpower<- -optint[[2]]
result<-round(c(r,ga1,totalpower),3)
names(result)<-c("r","gamma1","Power")
result
}
```

E Tables

The following tables show simulation results for the different selection procedures for constructing a linear score discussed in Section 4.3. Different FDR's for using the protected procedure, different numbers k of "best" markers for using the optimistic procedure and different boundaries γ when using the forward logistic regression as selection method are investigated.

The average number of selected true null hypothesis ($m_0^s \pm$SD), the average number of selected alternatives ($m_e^s \pm$SD) and the average future AUC ($\pm$SD) conditional that at least one marker was identified for future prediction are given in the tables for the different scenarios. The proportion of samples among all simulations steps where at least one marker has been identified for future prediction, $\hat{p}_s$, is only given in the tables concerning the logistic regression and the protected approach, since in the optimistic approach $\hat{p}_s \equiv 1$. The number of alternatives is assumed to be $m_e = 10$ and 60 among $m = 1000$ and 6000 tested hypotheses. The sample size is set to $n = 50$, 100 and 500 per group.

E.1 Simulation results for the optimistic approach

Table E.1: **Simulation results for the optimistic approach:** $m_e = 10$, $\Delta = 0.81$, $n = 50$ per group, $m = 1000$ and 6000.

	$m = 1000$						$m = 6000$					
best	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
2	0.00	(0.04)	2.00	(0.04)	0.790	(0.005)	0.01	(0.11)	1.99	(0.11)	0.789	(0.014)
5	0.04	(0.20)	4.96	(0.20)	0.897	(0.010)	0.20	(0.45)	4.80	(0.45)	0.889	(0.023)
8	0.44	(0.63)	7.56	(0.63)	0.936	(0.015)	1.14	(0.95)	6.86	(0.95)	0.917	(0.028)
10	1.39	(0.89)	8.61	(0.89)	0.942	(0.016)	2.40	(1.12)	7.60	(1.12)	0.919	(0.028)
15	5.64	(0.75)	9.36	(0.75)	0.935	(0.014)	6.59	(1.08)	8.41	(1.08)	0.910	(0.026)
20	10.41	(0.63)	9.59	(0.63)	0.926	(0.013)	11.25	(1.01)	8.75	(1.01)	0.898	(0.025)
25	15.30	(0.53)	9.71	(0.53)	0.918	(0.012)	16.03	(0.94)	8.97	(0.94)	0.887	(0.024)
30	20.23	(0.47)	9.78	(0.47)	0.910	(0.012)	20.88	(0.88)	9.12	(0.88)	0.876	(0.023)
35	25.18	(0.43)	9.82	(0.43)	0.903	(0.013)	25.78	(0.84)	9.22	(0.84)	0.867	(0.023)
40	30.15	(0.39)	9.85	(0.39)	0.897	(0.013)	30.70	(0.80)	9.30	(0.80)	0.859	(0.022)
50	40.11	(0.33)	9.89	(0.33)	0.887	(0.013)	40.58	(0.74)	9.42	(0.74)	0.844	(0.021)
60	50.08	(0.28)	9.92	(0.28)	0.878	(0.013)	50.49	(0.68)	9.51	(0.68)	0.832	(0.021)
100	90.03	(0.18)	9.97	(0.18)	0.855	(0.014)	90.30	(0.54)	9.70	(0.54)	0.797	(0.019)
1000	990.00	(0.00)	10.00	(0.00)	0.752	(0.015)	990.01	(0.11)	9.99	(0.11)	0.675	(0.012)

Table E.2: **Simulation results for the optimistic approach:** $m_e = 10$, $\Delta = 0.81$, $n = 100$ per group, $m = 1000$ and 6000.

	$m = 1000$						$m = 6000$					
best	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
2	0.00	(0.01)	2.00	(0.01)	0.791	(0.001)	0.00	(0.01)	2.00	(0.01)	0.791	(0.001)
5	0.00	(0.01)	5.00	(0.01)	0.899	(0.001)	0.00	(0.02)	5.00	(0.02)	0.899	(0.001)
8	0.00	(0.05)	8.00	(0.05)	0.946	(0.001)	0.01	(0.10)	7.99	(0.10)	0.946	(0.002)
10	0.09	(0.29)	9.91	(0.29)	0.962	(0.003)	0.24	(0.46)	9.76	(0.46)	0.960	(0.006)
15	5.01	(0.09)	9.99	(0.09)	0.954	(0.002)	5.05	(0.21)	9.95	(0.21)	0.950	(0.004)
20	10.00	(0.09)	10.00	(0.09)	0.948	(0.003)	10.02	(0.16)	9.97	(0.16)	0.941	(0.004)
25	15.00	(0.05)	10.00	(0.05)	0.942	(0.003)	15.02	(0.13)	9.98	(0.13)	0.932	(0.005)
30	20.00	(0.04)	10.00	(0.04)	0.938	(0.004)	20.01	(0.11)	9.99	(0.11)	0.925	(0.005)
35	25.00	(0.03)	10.00	(0.03)	0.934	(0.004)	25.01	(0.10)	9.99	(0.10)	0.918	(0.006)
40	30.00	(0.02)	10.00	(0.02)	0.930	(0.005)	30.01	(0.08)	9.99	(0.08)	0.911	(0.006)
50	40.00	(0.02)	10.00	(0.02)	0.923	(0.005)	40.01	(0.07)	10.00	(0.07)	0.900	(0.007)
60	50.00	(0.01)	10.00	(0.01)	0.918	(0.006)	50.00	(0.06)	10.00	(0.06)	0.890	(0.008)
100	90.00	(0.01)	10.00	(0.01)	0.902	(0.007)	90.00	(0.04)	10.00	(0.04)	0.860	(0.009)
1000	990.00	(0.00)	10.00	(0.00)	0.816	(0.011)	990.00	(0.00)	10.00	(0.00)	0.733	(0.010)

Table E.3: **Simulation results for the optimistic approach:** $m_e = 10$, $\Delta = 0.81$, $n = 500$ per group, $m = 1000$ and 6000.

	$m = 1000$						$m = 6000$					
best	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
10	0	(0)	10	(0)	0.965	(0.000)	0	(0)	10	(0)	0.965	(0.000)
15	5	(0)	10	(0)	0.963	(0.000)	5	(0)	10	(0)	0.962	(0.000)
20	10	(0)	10	(0)	0.962	(0.000)	10	(0)	10	(0)	0.960	(0.000)
25	15	(0)	10	(0)	0.961	(0.001)	15	(0)	10	(0)	0.958	(0.001)
30	20	(0)	10	(0)	0.960	(0.001)	20	(0)	10	(0)	0.957	(0.001)
35	25	(0)	10	(0)	0.959	(0.001)	25	(0)	10	(0)	0.955	(0.001)
40	30	(0)	10	(0)	0.958	(0.001)	30	(0)	10	(0)	0.954	(0.001)
50	40	(0)	10	(0)	0.957	(0.001)	40	(0)	10	(0)	0.951	(0.001)
60	50	(0)	10	(0)	0.955	(0.001)	50	(0)	10	(0)	0.949	(0.001)
100	90	(0)	10	(0)	0.952	(0.001)	90	(0)	10	(0)	0.940	(0.002)
1000	990	(0)	10	(0)	0.923	(0.003)	990	(0)	10	(0)	0.875	(0.004)

Table E.4: **Simulation results for the optimistic approach:** $m_e = 60$, $\Delta = 0.81$, $n = 50$ per group, $m = 1000$ and 6000.

	$m = 1000$						$m = 6000$					
best	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
5	0.87	(0.86)	4.13	(0.86)	0.667	(0.032)	2.39	(1.14)	2.61	(1.14)	0.608	(0.045)
10	2.48	(1.39)	7.52	(1.39)	0.711	(0.034)	5.67	(1.57)	4.33	(1.57)	0.627	(0.044)
50	26.14	(3.12)	23.86	(3.12)	0.795	(0.029)	37.94	(2.80)	12.06	(2.80)	0.662	(0.035)
60	33.54	(3.27)	26.46	(3.27)	0.800	(0.028)	46.63	(2.94)	13.37	(2.94)	0.665	(0.034)
100	65.73	(3.49)	34.27	(3.49)	0.810	(0.024)	82.48	(3.30)	17.52	(3.30)	0.670	(0.030)
150	109.36	(3.39)	40.64	(3.39)	0.813	(0.020)	128.59	(3.55)	21.41	(3.55)	0.672	(0.027)
200	154.94	(3.18)	45.06	(3.18)	0.814	(0.018)	175.53	(3.66)	24.47	(3.66)	0.673	(0.024)
250	201.66	(2.98)	48.34	(2.98)	0.813	(0.017)	222.98	(3.75)	27.02	(3.75)	0.673	(0.023)
300	249.16	(2.70)	50.84	(2.70)	0.813	(0.016)	270.80	(3.77)	29.20	(3.77)	0.673	(0.021)
350	297.19	(2.47)	52.81	(2.47)	0.813	(0.015)	318.91	(3.78)	31.09	(3.78)	0.672	(0.020)
400	345.64	(2.23)	54.36	(2.23)	0.813	(0.015)	367.23	(3.80)	32.77	(3.80)	0.672	(0.019)
450	394.40	(2.00)	55.60	(2.00)	0.813	(0.015)	415.72	(3.77)	34.28	(3.77)	0.671	(0.018)
500	443.41	(1.78)	56.59	(1.78)	0.814	(0.015)	464.34	(3.74)	35.66	(3.74)	0.670	(0.018)
550	492.60	(1.58)	57.40	(1.58)	0.814	(0.015)	513.10	(3.69)	36.90	(3.69)	0.670	(0.017)
600	541.94	(1.38)	58.06	(1.38)	0.813	(0.015)	561.96	(3.66)	38.04	(3.66)	0.669	(0.016)
650	591.41	(1.18)	58.59	(1.18)	0.812	(0.015)	610.91	(3.62)	39.09	(3.62)	0.669	(0.016)
700	640.99	(1.00)	59.01	(1.00)	0.811	(0.015)	659.93	(3.59)	40.07	(3.59)	0.668	(0.016)
750	690.66	(0.81)	59.34	(0.81)	0.808	(0.015)	709.03	(3.55)	40.97	(3.55)	0.667	(0.015)
800	740.42	(0.65)	59.58	(0.65)	0.803	(0.015)	758.16	(3.51)	41.84	(3.51)	0.667	(0.015)
850	790.24	(0.49)	59.76	(0.49)	0.796	(0.015)	807.36	(3.47)	42.64	(3.47)	0.667	(0.014)
900	840.11	(0.34)	59.89	(0.34)	0.787	(0.015)	856.61	(3.43)	43.39	(3.43)	0.666	(0.014)
950	890.03	(0.18)	59.97	(0.18)	0.774	(0.015)	905.88	(3.37)	44.12	(3.37)	0.666	(0.014)
1000	940.00	(0.00)	60.00	(0.00)	0.752	(0.015)	955.21	(3.32)	44.79	(3.32)	0.665	(0.014)

Table E.5: **Simulation results for the optimistic approach:** $m_e = 60$, $\Delta = 0.81$, $n = 100$ per group, $m = 1000$ and 6000.

	$m = 1000$						$m = 6000$					
best	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
5	0.11	(0.33)	4.89	(0.33)	0.695	(0.012)	0.52	(0.70)	4.48	(0.70)	0.680	(0.026)
10	0.44	(0.66)	9.56	(0.66)	0.759	(0.015)	1.84	(1.26)	8.16	(1.26)	0.727	(0.030)
50	15.25	(2.81)	34.75	(2.81)	0.881	(0.017)	26.48	(3.17)	23.52	(3.17)	0.793	(0.030)
60	21.83	(2.98)	38.17	(2.98)	0.886	(0.016)	34.37	(3.32)	25.63	(3.32)	0.794	(0.029)
100	53.42	(2.92)	46.58	(2.92)	0.888	(0.013)	68.36	(3.56)	31.64	(3.56)	0.792	(0.025)
150	98.35	(2.56)	51.66	(2.56)	0.883	(0.011)	113.61	(3.58)	36.39	(3.58)	0.785	(0.022)
200	145.56	(2.21)	54.44	(2.21)	0.878	(0.010)	160.35	(3.51)	39.65	(3.51)	0.778	(0.020)
250	193.83	(1.87)	56.17	(1.87)	0.874	(0.010)	160.35	(3.44)	39.65	(3.44)	0.772	(0.019)
300	242.70	(1.60)	57.30	(1.60)	0.872	(0.010)	255.95	(3.33)	44.05	(3.33)	0.767	(0.017)
350	291.93	(1.36)	58.07	(1.36)	0.870	(0.009)	304.37	(3.25)	45.64	(3.25)	0.762	(0.017)
400	341.38	(1.16)	58.62	(1.16)	0.870	(0.009)	353.05	(3.15)	46.95	(3.15)	0.758	(0.016)
450	390.99	(0.99)	59.01	(0.99)	0.869	(0.009)	401.91	(3.05)	48.09	(3.05)	0.754	(0.015)
500	440.70	(0.84)	59.30	(0.84)	0.869	(0.009)	450.93	(2.96)	49.07	(2.96)	0.750	(0.015)
550	490.50	(0.71)	59.50	(0.71)	0.869	(0.009)	500.08	(2.87)	49.92	(2.87)	0.747	(0.014)
600	540.34	(0.58)	59.66	(0.58)	0.869	(0.009)	549.32	(2.79)	50.68	(2.79)	0.745	(0.014)
650	590.23	(0.48)	59.77	(0.48)	0.868	(0.009)	598.66	(2.71)	51.34	(2.71)	0.742	(0.013)
700	640.15	(0.40)	59.85	(0.40)	0.867	(0.009)	648.06	(2.62)	51.94	(2.62)	0.740	(0.013)
750	690.09	(0.30)	59.91	(0.30)	0.865	(0.009)	697.51	(2.54)	52.49	(2.54)	0.738	(0.013)
800	740.05	(0.22)	59.95	(0.22)	0.862	(0.010)	747.01	(2.46)	52.99	(2.46)	0.736	(0.012)
850	790.03	(0.16)	59.98	(0.16)	0.856	(0.010)	796.56	(2.39)	53.43	(2.39)	0.734	(0.012)
900	840.01	(0.10)	59.99	(0.10)	0.849	(0.010)	846.14	(2.34)	53.86	(2.34)	0.732	(0.012)
950	890.00	(0.06)	60.00	(0.06)	0.837	(0.010)	895.77	(2.27)	54.23	(2.27)	0.731	(0.012)
1000	940.00	(0.00)	60.00	(0.00)	0.816	(0.011)	945.42	(2.22)	54.58	(2.22)	0.730	(0.012)

Table E.6: **Simulation results for the optimistic approach:** $m_e = 60$, $\Delta = 0.81$, $n = 500$ per group, $m = 1000$ and 6000.

	$m = 1000$						$m = 6000$					
best	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
5	0.00	(0)	5.00	(0)	0.699	(0.000)	0.00	(0.00)	5.00	(0.00)	0.699	(0.000)
10	0.00	(0)	10.00	(0)	0.770	(0.000)	0.00	(0.00)	10.00	(0.00)	0.770	(0.000)
50	0.01	(0.10)	49.99	(0.10)	0.949	(0.000)	0.07	(0.27)	49.93	(0.27)	0.949	(0.001)
60	0.89	(0.74)	59.11	(0.74)	0.961	(0.001)	1.97	(1.05)	58.03	(1.05)	0.959	(0.002)
100	40.01	(0.11)	59.99	(0.11)	0.955	(0.001)	40.17	(0.41)	59.83	(0.41)	0.949	(0.001)
150	90.00	(0.06)	60.00	(0.06)	0.950	(0.001)	90.06	(0.25)	59.94	(0.25)	0.938	(0.002)
200	140.00	(0.03)	60.00	(0.03)	0.947	(0.001)	140.04	(0.19)	59.97	(0.19)	0.929	(0.002)
250	190.00	(0.02)	60.00	(0.02)	0.945	(0.002)	190.02	(0.15)	59.98	(0.15)	0.922	(0.002)
300	240.00	(0.01)	60.00	(0.01)	0.944	(0.002)	240.02	(0.13)	59.98	(0.13)	0.916	(0.003)
350	290.00	(0.01)	60.00	(0.01)	0.943	(0.002)	290.01	(0.11)	59.99	(0.11)	0.910	(0.003)
400	340.00	(0)	60.00	(0)	0.943	(0.002)	340.01	(0.09)	59.99	(0.09)	0.906	(0.003)
450	390.00	(0)	60.00	(0)	0.943	(0.002)	390.01	(0.08)	59.99	(0.08)	0.902	(0.003)
500	440.00	(0)	60.00	(0)	0.942	(0.002)	440.01	(0.07)	60.00	(0.07)	0.898	(0.003)
550	490.00	(0)	60.00	(0)	0.942	(0.002)	490.00	(0.06)	60.00	(0.06)	0.895	(0.003)
600	540.00	(0)	60.00	(0)	0.942	(0.002)	540.00	(0.06)	60.00	(0.06)	0.892	(0.004)
650	590.00	(0)	60.00	(0)	0.942	(0.002)	590.00	(0.05)	60.00	(0.05)	0.889	(0.004)
700	640.00	(0)	60.00	(0)	0.942	(0.002)	640.00	(0.04)	60.00	(0.04)	0.886	(0.004)
750	690.00	(0)	60.00	(0)	0.941	(0.002)	690.00	(0.04)	60.00	(0.04)	0.884	(0.004)
800	740.00	(0)	60.00	(0)	0.940	(0.002)	740.00	(0.04)	60.00	(0.04)	0.882	(0.004)
850	790.00	(0)	60.00	(0)	0.939	(0.002)	790.00	(0.04)	60.00	(0.04)	0.880	(0.004)
900	840.00	(0)	60.00	(0)	0.936	(0.002)	840.00	(0.04)	60.00	(0.04)	0.878	(0.004)
950	890.00	(0)	60.00	(0)	0.932	(0.002)	890.00	(0.04)	60.00	(0.04)	0.876	(0.004)
1000	940.00	(0)	60.00	(0)	0.923	(0.003)	940.00	(0.04)	60.00	(0.04)	0.875	(0.004)

E.2 Simulation results for the forward logistic regression

Table E.7: **Simulation results for selection using forward logistic regression:** $m = 1000$, $n = 50$ and 100 per group, $m_e = 10$ and 60.

		$n = 50$				$n = 100$			
m_e	γ	$\hat{p}_s$	m_0^s (SD)	m_e^s (SD)	AUC (SD)	$\hat{p}_s$	m_0^s (SD)	m_e^s (SD)	AUC (SD)
10	0.00005	0.988	0.03 (0.18)	1.85 (0.70)	0.773 (0.045)	1	0.06 (0.24)	4.41 (0.86)	0.879 (0.024)
	0.0002	0.999	0.18 (0.46)	2.47 (0.73)	0.803 (0.041)	1	0.24 (0.53)	5.10 (0.91)	0.894 (0.022)
	0.0005	1	0.64 (0.94)	2.88 (0.82)	0.812 (0.043)	1	0.73 (1.10)	5.64 (0.93)	0.900 (0.021)
	0.0025	1	3.61 (1.56)	3.52 (0.87)	0.799 (0.049)	1	5.17 (2.50)	6.46 (0.98)	0.882 (0.029)
	0.0043	1	4.14 (1.33)	3.58 (0.91)	0.796 (0.050)	1	6.36 (1.88)	6.50 (0.95)	0.877 (0.027)
	0.05	1	4.22 (1.24)	3.63 (0.88)	0.797 (0.049)	1	7.01 (1.75)	6.57 (0.95)	0.876 (0.027)
60	0.00005	0.260	0.10 (0.32)	1.01 (0.41)	0.588 (0.027)	0.863	0.04 (0.19)	1.88 (0.90)	0.621 (0.029)
	0.0001	0.412	0.18 (0.41)	1.06 (0.55)	0.587 (0.033)	0.959	0.11 (0.34)	2.37 (1.12)	0.634 (0.033)
	0.0006	0.877	0.83 (1.08)	1.71 (1.03)	0.599 (0.042)	1	0.92 (1.20)	4.79 (1.60)	0.679 (0.032)
	0.0073	1	6.07 (1.56)	3.56 (1.33)	0.613 (0.040)	1	9.77 (2.18)	8.99 (1.69)	0.696 (0.032)
	0.0156	1	6.09 (1.56)	3.56 (1.32)	0.613 (0.040)	1	9.89 (2.17)	9.00 (1.69)	0.696 (0.032)
	0.05	1	6.13 (1.57)	3.53 (1.30)	0.612 (0.039)	1	10.28 (2.29)	9.02 (1.69)	0.696 (0.032)

E.3 Simulation results for the protected approach

Table E.8: **Simulation results for the protected approach:** $m_e = 10$, $\Delta = 0.81$, $n = 50$ per group, $m = 1000$ and 6000.

	$m = 1000$							$m = 6000$						
FDR	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
0.00005	0.670	0.00	(0.00)	1.78	(1.04)	0.765	(0.057)	0.439	0.00	(0.02)	1.43	(0.71)	0.746	(0.044)
0.0001	0.751	0.00	(0.00)	2.03	(1.19)	0.777	(0.062)	0.523	0.00	(0.02)	1.56	(0.82)	0.754	(0.049)
0.0005	0.903	0.00	(0.04)	2.99	(1.54)	0.822	(0.064)	0.727	0.00	(0.04)	1.99	(1.13)	0.776	(0.060)
0.001	0.944	0.00	(0.06)	3.51	(1.68)	0.842	(0.063)	0.806	0.00	(0.06)	2.29	(1.28)	0.791	(0.063)
0.005	0.991	0.03	(0.16)	4.97	(1.88)	0.885	(0.053)	0.932	0.02	(0.14)	3.23	(1.70)	0.829	(0.069)
0.01	0.996	0.07	(0.26)	5.73	(1.87)	0.903	(0.045)	0.965	0.05	(0.23)	3.84	(1.83)	0.850	(0.066)
0.05	1	0.44	(0.71)	7.53	(1.56)	0.933	(0.025)	0.996	0.32	(0.60)	5.58	(1.84)	0.897	(0.046)
0.10	1	1.00	(1.14)	8.23	(1.36)	0.939	(0.019)	0.999	0.76	(1.02)	6.40	(1.76)	0.910	(0.037)
0.15	1	1.66	(1.59)	8.59	(1.21)	0.941	(0.017)	0.999	1.31	(1.40)	6.92	(1.69)	0.915	(0.033)
0.20	1	2.42	(2.02)	8.85	(1.09)	0.941	(0.015)	1	1.96	(1.86)	7.28	(1.62)	0.917	(0.030)
0.25	1	3.31	(2.48)	9.04	(1.01)	0.940	(0.015)	1	2.75	(2.31)	7.59	(1.52)	0.918	(0.028)
0.30	1	4.33	(3.09)	9.17	(0.94)	0.938	(0.015)	1	3.69	(2.94)	7.83	(1.48)	0.916	(0.028)
0.35	1	5.59	(3.82)	9.29	(0.88)	0.935	(0.016)	1	4.77	(3.67)	8.02	(1.44)	0.914	(0.027)
0.40	1	7.03	(4.63)	9.39	(0.81)	0.933	(0.016)	1	6.15	(4.45)	8.22	(1.37)	0.911	(0.027)
0.45	1	8.73	(5.67)	9.47	(0.76)	0.930	(0.017)	1	7.66	(5.46)	8.38	(1.30)	0.908	(0.028)
0.50	1	10.85	(7.02)	9.55	(0.70)	0.926	(0.018)	1	9.61	(6.66)	8.54	(1.24)	0.903	(0.028)
0.55	1	13.58	(8.95)	9.61	(0.65)	0.921	(0.020)	1	12.18	(8.44)	8.69	(1.18)	0.898	(0.030)
0.60	1	16.99	(11.39)	9.68	(0.59)	0.916	(0.022)	1	15.32	(10.75)	8.82	(1.13)	0.891	(0.032)
0.65	1	21.49	(14.86)	9.73	(0.54)	0.910	(0.024)	1	19.52	(14.07)	8.95	(1.08)	0.883	(0.034)
0.70	1	28.24	(20.54)	9.78	(0.50)	0.903	(0.026)	1	25.73	(19.02)	9.09	(1.01)	0.873	(0.038)
0.75	1	38.39	(30.20)	9.83	(0.43)	0.894	(0.029)	1	34.52	(26.63)	9.22	(0.93)	0.861	(0.042)
0.80	1	55.55	(50.51)	9.87	(0.38)	0.883	(0.033)	1	49.46	(40.25)	9.36	(0.86)	0.845	(0.047)
0.85	1	93.08	(100.28)	9.91	(0.31)	0.868	(0.036)	1	77.57	(70.16)	9.50	(0.76)	0.824	(0.053)
0.90	1	178.55	(191.89)	9.95	(0.24)	0.851	(0.038)	1	150.22	(164.52)	9.64	(0.65)	0.793	(0.061)
0.95	1	374.82	(324.86)	9.98	(0.16)	0.827	(0.042)	1	540.96	(772.33)	9.81	(0.50)	0.744	(0.067)
0.99	1	619.80	(373.86)	9.99	(0.10)	0.796	(0.047)	1	2696.55	(2302.57)	9.94	(0.28)	0.678	(0.062)

Table E.9: **Simulation results for the protected approach:** $m_e = 10$, $\Delta = 0.81$, $n = 100$ per group, $m = 1000$ and 6000.

	$m = 1000$							$m = 6000$						
FDR	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
0.00005	1	0.00	(0.02)	7.83	(1.40)	0.942	(0.020)	1	0.00	(0.01)	6.60	(1.67)	0.923	(0.031)
0.0001	1	0.00	(0.03)	8.24	(1.27)	0.947	(0.016)	1	0.00	(0.02)	7.10	(1.59)	0.931	(0.026)
0.0005	1	0.00	(0.07)	9.00	(0.97)	0.955	(0.010)	1	0.00	(0.06)	8.13	(1.32)	0.945	(0.017)
0.001	1	0.01	(0.10)	9.24	(0.85)	0.957	(0.008)	1	0.01	(0.09)	8.50	(1.19)	0.950	(0.014)
0.005	1	0.05	(0.22)	9.65	(0.59)	0.960	(0.005)	1	0.04	(0.21)	9.18	(0.89)	0.956	(0.009)
0.01	1	0.11	(0.34)	9.77	(0.47)	0.961	(0.004)	1	0.10	(0.31)	9.40	(0.77)	0.958	(0.007)
0.05	1	0.57	(0.79)	9.93	(0.26)	0.961	(0.003)	1	0.55	(0.80)	9.76	(0.50)	0.959	(0.005)
0.10	1	1.19	(1.22)	9.96	(0.19)	0.960	(0.003)	1	1.20	(1.22)	9.85	(0.39)	0.958	(0.005)
0.15	1	1.87	(1.64)	9.98	(0.15)	0.959	(0.004)	1	1.86	(1.60)	9.89	(0.33)	0.957	(0.006)
0.20	1	2.65	(2.07)	9.99	(0.12)	0.958	(0.004)	1	2.71	(2.05)	9.92	(0.29)	0.955	(0.006)
0.25	1	3.56	(2.55)	9.99	(0.10)	0.956	(0.005)	1	3.56	(2.54)	9.93	(0.26)	0.953	(0.007)
0.30	1	4.57	(3.10)	9.99	(0.09)	0.955	(0.006)	1	4.61	(3.10)	9.94	(0.24)	0.951	(0.008)
0.35	1	5.83	(3.75)	9.99	(0.08)	0.953	(0.006)	1	5.85	(3.77)	9.95	(0.21)	0.948	(0.009)
0.40	1	7.29	(4.60)	9.99	(0.07)	0.951	(0.007)	1	7.27	(4.54)	9.96	(0.19)	0.946	(0.010)
0.45	1	9.06	(5.68)	10.00	(0.06)	0.949	(0.008)	1	8.93	(5.55)	9.97	(0.17)	0.943	(0.012)
0.50	1	11.20	(6.98)	10.00	(0.06)	0.947	(0.009)	1	11.03	(6.79)	9.97	(0.16)	0.939	(0.013)
0.55	1	13.90	(8.85)	10.00	(0.05)	0.944	(0.010)	1	13.75	(8.53)	9.98	(0.15)	0.935	(0.015)
0.60	1	17.29	(11.33)	10.00	(0.04)	0.941	(0.012)	1	17.06	(10.81)	9.98	(0.13)	0.930	(0.017)
0.65	1	21.87	(14.96)	10.00	(0.03)	0.937	(0.013)	1	21.44	(13.87)	9.99	(0.12)	0.924	(0.020)
0.70	1	28.60	(20.26)	10.00	(0.02)	0.932	(0.015)	1	27.82	(18.80)	9.99	(0.11)	0.916	(0.024)
0.75	1	38.50	(29.76)	10.00	(0.02)	0.927	(0.018)	1	37.07	(26.42)	9.99	(0.09)	0.907	(0.028)
0.80	1	55.75	(49.87)	10.00	(0.02)	0.920	(0.021)	1	51.69	(39.37)	9.99	(0.08)	0.895	(0.033)
0.85	1	91.73	(96.83)	10.00	(0.01)	0.910	(0.024)	1	80.12	(70.03)	10.00	(0.06)	0.877	(0.041)
0.90	1	176.60	(188.10)	10.00	(0.01)	0.898	(0.026)	1	152.06	(168.62)	10.00	(0.05)	0.851	(0.052)
0.95	1	377.32	(324.87)	10.00	(0.01)	0.879	(0.032)	1	550.63	(798.96)	10.00	(0.03)	0.804	(0.064)
0.99	1	626.10	(374.97)	10.00	(0.00)	0.854	(0.038)	1	2696.01	(2315.11)	10.00	(0.01)	0.733	(0.065)

Table E.10: **Simulation results for the protected approach:** $m_e = 10$, $\Delta = 0.81$, $n = 500$ per group, $m = 1000$ and 6000.

	$m = 1000$							$m = 6000$						
FDR	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	$\hat{p}_s$	(SD)	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
0.00005	1	0.00	(0.02)	10	(0)	0.965	(0.000)	1	0.00	(0.03)	10	(0)	0.965	(0.000)
0.0001	1	0.00	(0.04)	10	(0)	0.965	(0.000)	1	0.00	(0.03)	10	(0)	0.965	(0.000)
0.0005	1	0.01	(0.08)	10	(0)	0.965	(0.000)	1	0.01	(0.07)	10	(0)	0.965	(0.000)
0.001	1	0.01	(0.10)	10	(0)	0.965	(0.000)	1	0.01	(0.11)	10	(0)	0.965	(0.000)
0.005	1	0.06	(0.24)	10	(0)	0.965	(0.000)	1	0.05	(0.23)	10	(0)	0.965	(0.000)
0.01	1	0.11	(0.34)	10	(0)	0.965	(0.000)	1	0.11	(0.33)	10	(0)	0.965	(0.000)
0.05	1	0.58	(0.80)	10	(0)	0.964	(0.000)	1	0.58	(0.80)	10	(0)	0.964	(0.001)
0.10	1	1.23	(1.23)	10	(0)	0.964	(0.001)	1	1.19	(1.21)	10	(0)	0.964	(0.001)
0.15	1	1.91	(1.65)	10	(0)	0.964	(0.001)	1	1.87	(1.61)	10	(0)	0.964	(0.001)
0.20	1	2.73	(2.10)	10	(0)	0.964	(0.001)	1	2.71	(2.06)	10	(0)	0.963	(0.001)
0.25	1	3.66	(2.57)	10	(0)	0.963	(0.001)	1	3.55	(2.54)	10	(0)	0.963	(0.001)
0.30	1	4.69	(3.14)	10	(0)	0.963	(0.001)	1	4.61	(3.11)	10	(0)	0.962	(0.002)
0.35	1	5.97	(3.82)	10	(0)	0.963	(0.001)	1	5.88	(3.79)	10	(0)	0.962	(0.002)
0.40	1	7.41	(4.64)	10	(0)	0.962	(0.001)	1	7.30	(4.57)	10	(0)	0.961	(0.002)
0.45	1	9.19	(5.73)	10	(0)	0.962	(0.002)	1	8.91	(5.55)	10	(0)	0.961	(0.002)
0.50	1	11.36	(7.02)	10	(0)	0.961	(0.002)	1	11.02	(6.86)	10	(0)	0.960	(0.003)
0.55	1	14.12	(8.85)	10	(0)	0.961	(0.002)	1	13.73	(8.54)	10	(0)	0.959	(0.003)
0.60	1	17.54	(11.09)	10	(0)	0.960	(0.002)	1	16.92	(10.83)	10	(0)	0.958	(0.004)
0.65	1	22.06	(14.65)	10	(0)	0.959	(0.003)	1	21.34	(13.99)	10	(0)	0.957	(0.005)
0.70	1	28.86	(19.97)	10	(0)	0.958	(0.003)	1	27.59	(18.72)	10	(0)	0.955	(0.006)
0.75	1	39.01	(30.09)	10	(0)	0.957	(0.004)	1	36.61	(26.35)	10	(0)	0.952	(0.007)
0.80	1	56.25	(51.59)	10	(0)	0.955	(0.005)	1	51.61	(40.22)	10	(0)	0.949	(0.009)
0.85	1	92.14	(96.67)	10	(0)	0.953	(0.006)	1	79.98	(70.47)	10	(0)	0.943	(0.013)
0.90	1	178.29	(188.33)	10	(0)	0.950	(0.006)	1	150.52	(167.16)	10	(0)	0.934	(0.019)
0.95	1	376.90	(324.15)	10	(0)	0.945	(0.009)	1	542.15	(772.02)	10	(0)	0.911	(0.031)
0.99	1	622.81	(372.95)	10	(0)	0.936	(0.012)	1	2665.36	(2298.34)	10	(0)	0.864	(0.045)

Table E.11: **Simulation results for the protected approach:** $m_e = 60$, $\Delta = 0.33$, $n = 50$ per group, $m = 1000$ and 6000.

	$m = 1000$							$m = 6000$						
FDR	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
0.00005	0.007	0.01	(0.12)	1.00	(0.17)	0.592	(0.012)	0.002	0.00	(0.00)	1.00	(0.00)	0.592	(0.000)
0.0001	0.012	0.01	(0.09)	1.00	(0.13)	0.592	(0.009)	0.003	0.06	(0.25)	0.97	(0.18)	0.589	(0.017)
0.0005	0.035	0.01	(0.09)	1.02	(0.19)	0.593	(0.011)	0.011	0.04	(0.19)	0.97	(0.16)	0.590	(0.015)
0.001	0.056	0.01	(0.12)	1.05	(0.28)	0.594	(0.015)	0.018	0.06	(0.23)	0.96	(0.25)	0.588	(0.021)
0.005	0.165	0.03	(0.18)	1.14	(0.49)	0.596	(0.021)	0.059	0.10	(0.30)	0.94	(0.34)	0.585	(0.027)
0.01	0.250	0.05	(0.22)	1.28	(0.68)	0.600	(0.026)	0.090	0.12	(0.33)	0.94	(0.40)	0.584	(0.030)
0.05	0.582	0.19	(0.48)	2.11	(1.65)	0.620	(0.045)	0.260	0.27	(0.50)	1.13	(0.73)	0.586	(0.039)
0.10	0.766	0.53	(0.90)	3.39	(2.59)	0.646	(0.056)	0.405	0.44	(0.68)	1.31	(0.95)	0.589	(0.043)
0.15	0.861	1.08	(1.48)	4.78	(3.44)	0.668	(0.061)	0.493	0.63	(0.94)	1.49	(1.22)	0.591	(0.046)
0.20	0.918	1.98	(2.24)	6.56	(4.21)	0.691	(0.063)	0.593	0.93	(1.21)	1.77	(1.39)	0.596	(0.048)
0.25	0.957	3.31	(3.25)	8.54	(4.94)	0.712	(0.062)	0.659	1.41	(1.71)	2.12	(1.71)	0.601	(0.050)
0.30	0.972	5.22	(4.65)	10.70	(5.68)	0.729	(0.062)	0.705	1.90	(2.33)	2.45	(2.05)	0.604	(0.052)
0.35	0.985	7.90	(6.38)	13.15	(6.26)	0.746	(0.057)	0.736	2.36	(3.09)	2.67	(2.38)	0.606	(0.054)
0.40	0.992	11.59	(8.52)	15.81	(6.77)	0.761	(0.053)	0.810	3.49	(4.12)	3.28	(2.68)	0.613	(0.054)
0.45	0.996	16.66	(11.33)	18.68	(7.16)	0.774	(0.048)	0.837	4.98	(5.69)	3.92	(3.12)	0.619	(0.054)
0.50	0.998	23.64	(15.09)	21.86	(7.51)	0.785	(0.042)	0.886	6.97	(7.58)	4.69	(3.49)	0.626	(0.053)
0.55	0.999	33.32	(20.07)	25.26	(7.73)	0.794	(0.037)	0.902	9.81	(10.40)	5.56	(4.02)	0.631	(0.053)
0.60	0.999	46.69	(26.79)	28.91	(7.88)	0.801	(0.032)	0.932	13.98	(14.28)	6.69	(4.54)	0.638	(0.051)
0.65	1	66.14	(36.85)	32.94	(7.96)	0.806	(0.028)	0.948	20.76	(20.55)	8.15	(5.24)	0.644	(0.050)
0.70	1	94.68	(52.52)	37.26	(7.98)	0.810	(0.024)	0.961	31.46	(30.29)	9.97	(6.05)	0.650	(0.048)
0.75	1	138.73	(78.29)	41.88	(7.78)	0.812	(0.021)	0.984	49.56	(45.89)	12.43	(6.93)	0.657	(0.043)
0.80	1	208.69	(118.72)	46.70	(7.43)	0.813	(0.018)	0.989	83.01	(73.21)	15.83	(7.95)	0.662	(0.038)
0.85	1	325.44	(174.29)	51.59	(6.60)	0.813	(0.017)	0.995	154.30	(132.88)	20.82	(9.32)	0.667	(0.033)
0.90	1	516.68	(241.35)	55.82	(5.07)	0.806	(0.023)	0.999	350.51	(315.54)	28.67	(11.04)	0.669	(0.026)
0.95	1	771.53	(237.68)	58.65	(3.06)	0.777	(0.032)	1	1199.01	(1094.75)	41.90	(12.43)	0.666	(0.019)
0.99	1	895.90	(145.48)	59.65	(1.52)	0.758	(0.024)	1	3943.52	(2083.72)	55.12	(8.33)	0.642	(0.025)

Table E.12: **Simulation results for the protected approach:** $m_e = 60, \Delta = 0.33$, $n = 100$ per group, $m = 1000$ and 6000.

	$m = 1000$							$m = 6000$						
FDR	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
0.00005	0.083	0.00	(0.00)	1.06	(0.25)	0.595	(0.009)	0.030	0.00	(0.06)	1.03	(0.18)	0.593	(0.008)
0.0001	0.121	0.00	(0.03)	1.10	(0.31)	0.596	(0.012)	0.046	0.00	(0.05)	1.04	(0.22)	0.594	(0.009)
0.0005	0.290	0.00	(0.05)	1.35	(0.67)	0.604	(0.022)	0.112	0.01	(0.08)	1.10	(0.35)	0.596	(0.014)
0.001	0.401	0.00	(0.07)	1.55	(0.88)	0.611	(0.027)	0.165	0.01	(0.09)	1.14	(0.44)	0.597	(0.017)
0.005	0.715	0.02	(0.16)	2.68	(1.81)	0.640	(0.045)	0.367	0.02	(0.13)	1.36	(0.77)	0.604	(0.026)
0.01	0.848	0.05	(0.24)	3.92	(2.56)	0.666	(0.053)	0.490	0.03	(0.18)	1.58	(1.05)	0.610	(0.032)
0.05	0.991	0.63	(0.91)	11.25	(4.92)	0.767	(0.056)	0.832	0.24	(0.53)	3.38	(2.26)	0.652	(0.050)
0.10	0.999	2.09	(1.83)	17.94	(5.68)	0.820	(0.043)	0.928	0.70	(1.05)	5.31	(3.28)	0.684	(0.057)
0.15	1	4.24	(2.83)	23.05	(5.87)	0.847	(0.034)	0.962	1.50	(1.70)	7.36	(4.01)	0.710	(0.058)
0.20	1	7.04	(3.96)	27.28	(5.86)	0.863	(0.028)	0.982	2.62	(2.54)	9.28	(4.61)	0.729	(0.057)
0.25	1	10.68	(5.29)	31.00	(5.73)	0.873	(0.023)	0.991	4.17	(3.52)	11.31	(4.96)	0.746	(0.052)
0.30	1	15.17	(6.76)	34.25	(5.57)	0.879	(0.020)	0.995	6.24	(4.75)	13.29	(5.33)	0.759	(0.049)
0.35	1	20.68	(8.57)	37.14	(5.38)	0.883	(0.018)	0.998	8.98	(6.24)	15.30	(5.70)	0.768	(0.047)
0.40	1	27.52	(10.91)	39.84	(5.19)	0.886	(0.016)	0.999	12.61	(8.09)	17.45	(5.85)	0.777	(0.042)
0.45	1	35.95	(13.64)	42.38	(5.00)	0.887	(0.015)	1	17.25	(10.41)	19.59	(6.00)	0.783	(0.038)
0.50	1	46.47	(17.03)	44.76	(4.74)	0.887	(0.014)	1	23.47	(13.43)	21.86	(6.10)	0.788	(0.034)
0.55	1	59.83	(21.73)	47.01	(4.47)	0.887	(0.013)	1	31.73	(17.32)	24.21	(6.21)	0.791	(0.031)
0.60	1	77.43	(28.31)	49.20	(4.17)	0.885	(0.013)	1	43.06	(22.91)	26.71	(6.37)	0.792	(0.029)
0.65	1	100.52	(37.61)	51.23	(3.85)	0.883	(0.013)	1	58.72	(30.48)	29.40	(6.41)	0.791	(0.027)
0.70	1	132.44	(52.35)	53.19	(3.51)	0.880	(0.012)	1	81.61	(41.23)	32.34	(6.41)	0.789	(0.025)
0.75	1	178.72	(75.52)	55.03	(3.11)	0.876	(0.012)	1	116.44	(58.93)	35.56	(6.41)	0.784	(0.024)
0.80	1	250.16	(112.85)	56.70	(2.67)	0.873	(0.011)	1	173.44	(89.35)	39.23	(6.40)	0.777	(0.023)
0.85	1	364.91	(166.31)	58.13	(2.12)	0.870	(0.011)	1	278.48	(152.02)	43.46	(6.29)	0.767	(0.023)
0.90	1	553.93	(232.88)	59.17	(1.43)	0.861	(0.020)	1	522.90	(331.75)	48.53	(6.01)	0.751	(0.023)
0.95	1	797.54	(216.71)	59.78	(0.73)	0.836	(0.026)	1	1424.64	(1085.72)	54.51	(5.06)	0.730	(0.021)
0.99	1	907.13	(123.80)	59.95	(0.35)	0.820	(0.018)	1	4114.42	(1983.12)	58.79	(2.64)	0.691	(0.031)

Table E.13: **Simulation results for the protected approach:** $m_e = 60$, $\Delta = 0.33$, $n = 500$ per group, $m = 1000$ and 6000.

	$m = 1000$							$m = 6000$						
FDR	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
0.00005	1	0.00	(0.04)	44.57	(3.74)	0.939	(0.008)	1	0.00	(0.04)	35.71	(4.35)	0.917	(0.013)
0.0001	1	0.00	(0.07)	47.50	(3.44)	0.945	(0.007)	1	0.00	(0.06)	39.20	(4.14)	0.926	(0.011)
0.0005	1	0.03	(0.16)	53.10	(2.60)	0.954	(0.004)	1	0.02	(0.16)	46.60	(3.50)	0.943	(0.007)
0.001	1	0.06	(0.24)	54.93	(2.23)	0.956	(0.003)	1	0.05	(0.23)	49.33	(3.16)	0.948	(0.006)
0.005	1	0.30	(0.54)	57.87	(1.48)	0.960	(0.002)	1	0.28	(0.53)	54.38	(2.35)	0.956	(0.004)
0.01	1	0.60	(0.78)	58.64	(1.17)	0.961	(0.002)	1	0.58	(0.78)	55.99	(1.99)	0.958	(0.003)
0.05	1	3.16	(1.88)	59.64	(0.60)	0.961	(0.001)	1	3.09	(1.86)	58.45	(1.25)	0.959	(0.002)
0.10	1	6.72	(2.91)	59.83	(0.41)	0.960	(0.001)	1	6.63	(2.88)	59.09	(0.96)	0.958	(0.002)
0.15	1	10.74	(3.91)	59.90	(0.31)	0.959	(0.001)	1	10.58	(3.82)	59.37	(0.80)	0.957	(0.002)
0.20	1	15.27	(4.94)	59.94	(0.24)	0.958	(0.001)	1	15.11	(4.87)	59.53	(0.69)	0.956	(0.002)
0.25	1	20.44	(6.17)	59.96	(0.20)	0.957	(0.002)	1	20.17	(6.01)	59.64	(0.61)	0.954	(0.002)
0.30	1	26.28	(7.56)	59.97	(0.17)	0.956	(0.002)	1	25.97	(7.35)	59.71	(0.54)	0.953	(0.003)
0.35	1	33.05	(9.25)	59.98	(0.14)	0.955	(0.002)	1	32.73	(8.90)	59.77	(0.48)	0.951	(0.003)
0.40	1	41.01	(11.15)	59.99	(0.12)	0.954	(0.002)	1	40.62	(10.80)	59.82	(0.43)	0.949	(0.003)
0.45	1	50.33	(13.70)	59.99	(0.10)	0.953	(0.002)	1	49.98	(13.12)	59.86	(0.38)	0.946	(0.004)
0.50	1	61.71	(16.94)	59.99	(0.08)	0.952	(0.002)	1	61.15	(16.03)	59.89	(0.34)	0.944	(0.004)
0.55	1	75.69	(21.24)	60.00	(0.07)	0.951	(0.003)	1	74.87	(19.65)	59.91	(0.30)	0.941	(0.005)
0.60	1	93.45	(27.23)	60.00	(0.05)	0.950	(0.003)	1	92.17	(24.70)	59.93	(0.26)	0.937	(0.006)
0.65	1	116.46	(36.12)	60.00	(0.04)	0.948	(0.003)	1	114.56	(31.72)	59.95	(0.23)	0.933	(0.006)
0.70	1	148.17	(50.20)	60.00	(0.03)	0.947	(0.003)	1	144.80	(41.75)	59.96	(0.19)	0.929	(0.007)
0.75	1	194.00	(73.44)	60.00	(0.02)	0.945	(0.003)	1	187.23	(57.81)	59.97	(0.16)	0.923	(0.009)
0.80	1	265.47	(111.80)	60.00	(0.01)	0.944	(0.003)	1	252.23	(85.34)	59.98	(0.13)	0.915	(0.010)
0.85	1	378.67	(164.87)	60.00	(0.00)	0.943	(0.003)	1	366.77	(145.24)	59.99	(0.10)	0.905	(0.013)
0.90	1	564.92	(230.99)	60.00	(0.00)	0.940	(0.006)	1	620.48	(320.96)	60.00	(0.06)	0.891	(0.015)
0.95	1	804.89	(212.78)	60.00	(0.00)	0.930	(0.009)	1	1526.99	(1058.90)	60.00	(0.03)	0.871	(0.015)
0.99	1	907.77	(121.22)	60.00	(0.00)	0.925	(0.006)	1	4192.20	(1932.61)	60.00	(0.01)	0.832	(0.032)

Table E.14: AUC_*=**0.8: Simulation results for the protected approach:** $m_e = 10$, $\Delta = 0.38$, $n = 50$ per group, $m = 1000$ and 6000.

	$m = 1000$							$m = 6000$						
FDR	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
0.00005	0.003	0.00	(0.00)	1.00	(0.00)	0.605	(0.000)	0.001	0.00	(0.00)	1.00	(0.00)	0.605	(0.000)
0.0001	0.004	0.00	(0.00)	1.00	(0.00)	0.605	(0.000)	0.001	0.00	(0.00)	1.00	(0.00)	0.605	(0.000)
0.0005	0.014	0.05	(0.22)	0.96	(0.23)	0.600	(0.023)	0.004	0.09	(0.29)	0.91	(0.29)	0.595	(0.031)
0.001	0.023	0.07	(0.25)	0.96	(0.28)	0.599	(0.026)	0.008	0.13	(0.34)	0.87	(0.34)	0.591	(0.035)
0.005	0.059	0.08	(0.27)	0.97	(0.34)	0.599	(0.029)	0.025	0.24	(0.44)	0.77	(0.42)	0.580	(0.044)
0.01	0.091	0.11	(0.32)	0.96	(0.41)	0.597	(0.034)	0.040	0.29	(0.46)	0.73	(0.46)	0.576	(0.047)
0.05	0.247	0.25	(0.48)	0.98	(0.60)	0.591	(0.043)	0.121	0.49	(0.60)	0.68	(0.57)	0.566	(0.052)
0.10	0.361	0.41	(0.66)	1.07	(0.76)	0.591	(0.047)	0.202	0.66	(0.73)	0.66	(0.62)	0.560	(0.053)
0.15	0.447	0.59	(0.85)	1.19	(0.86)	0.593	(0.048)	0.264	0.75	(0.86)	0.63	(0.64)	0.556	(0.053)
0.20	0.513	0.81	(1.15)	1.29	(0.98)	0.593	(0.050)	0.335	1.03	(1.04)	0.66	(0.69)	0.554	(0.052)
0.25	0.598	1.20	(1.47)	1.48	(1.08)	0.595	(0.050)	0.391	1.25	(1.33)	0.69	(0.74)	0.553	(0.052)
0.30	0.631	1.36	(1.80)	1.50	(1.17)	0.594	(0.051)	0.434	1.41	(1.62)	0.69	(0.76)	0.551	(0.051)
0.35	0.684	1.89	(2.36)	1.68	(1.28)	0.595	(0.051)	0.469	1.48	(1.89)	0.68	(0.77)	0.550	(0.051)
0.40	0.723	2.53	(3.12)	1.83	(1.40)	0.596	(0.051)	0.550	2.15	(2.44)	0.78	(0.84)	0.550	(0.049)
0.45	0.751	3.13	(4.06)	1.94	(1.53)	0.595	(0.051)	0.578	2.60	(3.21)	0.82	(0.89)	0.550	(0.048)
0.50	0.824	4.40	(5.15)	2.25	(1.62)	0.598	(0.049)	0.642	3.58	(4.13)	0.94	(0.97)	0.550	(0.047)
0.55	0.840	5.64	(6.93)	2.42	(1.73)	0.598	(0.048)	0.660	4.33	(5.44)	1.00	(1.04)	0.550	(0.046)
0.60	0.861	7.60	(9.30)	2.68	(1.87)	0.599	(0.047)	0.706	5.78	(7.10)	1.11	(1.12)	0.549	(0.044)
0.65	0.890	10.56	(12.87)	3.01	(1.99)	0.599	(0.045)	0.733	7.55	(9.72)	1.21	(1.21)	0.549	(0.043)
0.70	0.914	15.32	(18.69)	3.42	(2.13)	0.599	(0.043)	0.761	10.86	(14.27)	1.39	(1.34)	0.549	(0.041)
0.75	0.933	23.64	(29.71)	3.94	(2.32)	0.599	(0.040)	0.841	16.01	(21.07)	1.64	(1.47)	0.549	(0.038)
0.80	0.955	38.16	(50.40)	4.59	(2.44)	0.599	(0.037)	0.861	25.92	(35.55)	1.97	(1.68)	0.549	(0.036)
0.85	0.967	73.26	(101.07)	5.48	(2.65)	0.597	(0.033)	0.896	47.16	(65.66)	2.47	(1.94)	0.549	(0.032)
0.90	0.980	158.27	(196.52)	6.63	(2.72)	0.594	(0.029)	0.922	109.73	(162.14)	3.31	(2.34)	0.547	(0.029)
0.95	0.990	360.61	(334.80)	8.00	(2.47)	0.587	(0.025)	0.958	508.10	(816.06)	5.13	(2.95)	0.544	(0.022)
0.99	0.996	610.22	(384.01)	9.00	(1.89)	0.578	(0.022)	0.989	2685.61	(2328.15)	8.05	(2.70)	0.536	(0.015)

Table E.15: AUC_*=**0.8: Simulation results for the protected approach:** $m_e = 10$, $\Delta = 0.38$, $n = 100$ per group, $m = 1000$ and 6000.

	$m = 1000$							$m = 6000$						
FDR	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
0.00005	0.040	0.00	(0.00)	1.01	(0.12)	0.606	(0.005)	0.013	0.01	(0.09)	0.99	(0.09)	0.604	(0.009)
0.0001	0.057	0.00	(0.04)	1.03	(0.16)	0.606	(0.007)	0.020	0.02	(0.12)	1.00	(0.17)	0.604	(0.014)
0.0005	0.125	0.01	(0.09)	1.10	(0.35)	0.609	(0.016)	0.050	0.01	(0.12)	1.02	(0.24)	0.605	(0.015)
0.001	0.172	0.01	(0.10)	1.15	(0.43)	0.611	(0.018)	0.072	0.02	(0.13)	1.03	(0.27)	0.605	(0.017)
0.005	0.349	0.02	(0.14)	1.29	(0.63)	0.615	(0.026)	0.160	0.03	(0.18)	1.06	(0.36)	0.605	(0.022)
0.01	0.446	0.04	(0.19)	1.43	(0.77)	0.620	(0.030)	0.219	0.05	(0.22)	1.10	(0.46)	0.606	(0.026)
0.05	0.711	0.17	(0.45)	2.04	(1.29)	0.637	(0.042)	0.444	0.19	(0.45)	1.36	(0.79)	0.612	(0.039)
0.10	0.820	0.42	(0.76)	2.57	(1.57)	0.649	(0.047)	0.575	0.37	(0.67)	1.55	(0.98)	0.615	(0.044)
0.15	0.879	0.71	(1.08)	2.97	(1.70)	0.656	(0.047)	0.650	0.55	(0.91)	1.68	(1.14)	0.615	(0.047)
0.20	0.915	1.14	(1.49)	3.41	(1.85)	0.663	(0.048)	0.723	0.84	(1.18)	1.89	(1.22)	0.619	(0.048)
0.25	0.947	1.72	(1.96)	3.85	(1.89)	0.669	(0.046)	0.770	1.23	(1.59)	2.07	(1.36)	0.620	(0.049)
0.30	0.955	2.25	(2.56)	4.08	(2.01)	0.670	(0.046)	0.800	1.60	(2.04)	2.20	(1.46)	0.619	(0.050)
0.35	0.970	3.09	(3.21)	4.46	(2.03)	0.673	(0.045)	0.821	1.88	(2.56)	2.24	(1.54)	0.619	(0.050)
0.40	0.977	4.16	(4.04)	4.83	(2.06)	0.675	(0.043)	0.867	2.72	(3.24)	2.55	(1.58)	0.621	(0.048)
0.45	0.981	5.35	(5.05)	5.12	(2.13)	0.675	(0.042)	0.879	3.53	(4.13)	2.72	(1.68)	0.621	(0.048)
0.50	0.990	7.05	(6.29)	5.51	(2.05)	0.676	(0.039)	0.910	4.82	(5.20)	2.99	(1.74)	0.621	(0.046)
0.55	0.991	9.06	(8.15)	5.79	(2.10)	0.675	(0.038)	0.919	6.23	(6.93)	3.19	(1.85)	0.620	(0.046)
0.60	0.994	12.00	(10.66)	6.16	(2.09)	0.673	(0.036)	0.935	8.27	(8.97)	3.46	(1.91)	0.620	(0.044)
0.65	0.996	16.02	(14.30)	6.52	(2.06)	0.671	(0.035)	0.943	11.06	(12.09)	3.71	(2.01)	0.618	(0.043)
0.70	0.997	22.10	(19.95)	6.93	(1.99)	0.667	(0.033)	0.952	15.52	(16.85)	4.04	(2.10)	0.616	(0.041)
0.75	0.998	32.01	(30.28)	7.37	(1.93)	0.662	(0.032)	0.973	22.57	(24.64)	4.48	(2.13)	0.613	(0.039)
0.80	0.999	48.85	(50.29)	7.84	(1.82)	0.656	(0.031)	0.977	35.19	(39.44)	4.96	(2.21)	0.609	(0.036)
0.85	0.999	85.38	(97.95)	8.36	(1.69)	0.648	(0.030)	0.985	60.16	(70.05)	5.59	(2.26)	0.603	(0.034)
0.90	1	171.45	(192.78)	8.90	(1.47)	0.638	(0.029)	0.990	129.75	(175.32)	6.42	(2.32)	0.594	(0.032)
0.95	1	370.96	(328.20)	9.41	(1.14)	0.624	(0.028)	0.995	511.11	(782.50)	7.71	(2.23)	0.579	(0.029)
0.99	1	621.26	(377.16)	9.73	(0.79)	0.608	(0.027)	0.999	2650.89	(2321.50)	9.16	(1.61)	0.557	(0.023)

Table E.16: AUC_*=**0.8: Simulation results for the protected approach:** $m_e = 10$, $\Delta = 0.38$, $n = 500$ per group, $m = 1000$ and 6000.

	$m = 1000$							$m = 6000$						
FDR	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
0.00005	1	0.00	(0.02)	8.48	(1.19)	0.778	(0.017)	1	0.00	(0.01)	7.43	(1.52)	0.763	(0.024)
0.0001	1	0.00	(0.03)	8.79	(1.08)	0.782	(0.015)	1	0.00	(0.03)	7.87	(1.40)	0.769	(0.021)
0.0005	1	0.01	(0.08)	9.35	(0.80)	0.790	(0.010)	1	0.00	(0.07)	8.70	(1.12)	0.781	(0.016)
0.001	1	0.01	(0.11)	9.52	(0.69)	0.792	(0.009)	1	0.01	(0.10)	8.99	(1.00)	0.785	(0.014)
0.005	1	0.05	(0.23)	9.80	(0.45)	0.795	(0.006)	1	0.05	(0.22)	9.48	(0.72)	0.791	(0.009)
0.01	1	0.10	(0.32)	9.87	(0.36)	0.795	(0.005)	1	0.10	(0.31)	9.63	(0.61)	0.792	(0.008)
0.05	1	0.56	(0.77)	9.96	(0.20)	0.794	(0.004)	1	0.56	(0.78)	9.85	(0.39)	0.793	(0.006)
0.10	1	1.21	(1.20)	9.98	(0.15)	0.793	(0.005)	1	1.19	(1.20)	9.91	(0.30)	0.791	(0.006)
0.15	1	1.88	(1.60)	9.99	(0.12)	0.791	(0.005)	1	1.88	(1.60)	9.94	(0.25)	0.789	(0.007)
0.20	1	2.68	(2.05)	9.99	(0.11)	0.789	(0.006)	1	2.71	(2.05)	9.95	(0.22)	0.786	(0.008)
0.25	1	3.60	(2.54)	9.99	(0.10)	0.787	(0.007)	1	3.56	(2.54)	9.96	(0.20)	0.784	(0.009)
0.30	1	4.62	(3.12)	9.99	(0.08)	0.785	(0.007)	1	4.60	(3.07)	9.97	(0.18)	0.781	(0.010)
0.35	1	5.92	(3.86)	9.99	(0.07)	0.783	(0.008)	1	5.86	(3.77)	9.97	(0.16)	0.778	(0.011)
0.40	1	7.35	(4.62)	10.00	(0.06)	0.781	(0.009)	1	7.30	(4.60)	9.98	(0.14)	0.774	(0.012)
0.45	1	9.10	(5.72)	10.00	(0.06)	0.778	(0.010)	1	8.96	(5.62)	9.98	(0.13)	0.771	(0.013)
0.50	1	11.20	(6.93)	10.00	(0.05)	0.775	(0.011)	1	11.10	(6.84)	9.99	(0.11)	0.767	(0.014)
0.55	1	13.85	(8.60)	10.00	(0.04)	0.772	(0.012)	1	13.83	(8.60)	9.99	(0.10)	0.762	(0.016)
0.60	1	17.35	(11.15)	10.00	(0.04)	0.769	(0.013)	1	17.12	(10.81)	9.99	(0.09)	0.757	(0.018)
0.65	1	21.85	(14.70)	10.00	(0.03)	0.765	(0.014)	1	21.63	(14.01)	9.99	(0.08)	0.751	(0.020)
0.70	1	28.44	(20.07)	10.00	(0.02)	0.760	(0.016)	1	28.29	(19.23)	10.00	(0.07)	0.744	(0.022)
0.75	1	38.36	(30.04)	10.00	(0.02)	0.754	(0.018)	1	37.48	(26.89)	10.00	(0.06)	0.736	(0.025)
0.80	1	54.89	(48.81)	10.00	(0.02)	0.747	(0.020)	1	52.55	(40.19)	10.00	(0.06)	0.725	(0.028)
0.85	1	90.30	(94.44)	10.00	(0.01)	0.738	(0.022)	1	81.08	(69.67)	10.00	(0.05)	0.711	(0.032)
0.90	1	173.20	(185.62)	10.00	(0.01)	0.728	(0.023)	1	152.78	(159.82)	10.00	(0.03)	0.692	(0.037)
0.95	1	368.54	(321.81)	10.00	(0.00)	0.713	(0.025)	1	547.82	(774.15)	10.00	(0.01)	0.661	(0.042)
0.99	1	616.00	(373.52)	10.00	(0.00)	0.694	(0.028)	1	2716.95	(2319.46)	10.00	(0.00)	0.619	(0.040)

Table E.17: AUC_*=**0.8: Simulation results for the protected approach:** $m_e = 60$, $\Delta = 0.15$, $n = 50$ per group, $m = 1000$ and 6000.

	$m = 1000$							$m = 6000$						
FDR	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
0.00005	0.0002	1.00	(0.00)	0.00	(0.00)	0.500	(0.000)	0.0000	0.00	(NA)	0.00	(NA)	NA	(NA)
0.0001	0.0005	0.40	(0.55)	0.60	(0.55)	0.526	(0.024)	0.0001	1.00	(NA)	0.00	(NA)	0.500	(NA)
0.0005	0.002	0.33	(0.49)	0.67	(0.49)	0.529	(0.021)	0.0004	1.00	(0.00)	0.00	(0.00)	0.500	(0.000)
0.001	0.003	0.35	(0.49)	0.65	(0.49)	0.528	(0.021)	0.001	1.00	(0.00)	0.00	(0.00)	0.500	(0.000)
0.005	0.013	0.39	(0.49)	0.63	(0.51)	0.527	(0.021)	0.006	0.75	(0.44)	0.25	(0.44)	0.511	(0.019)
0.01	0.024	0.41	(0.50)	0.61	(0.51)	0.526	(0.021)	0.013	0.77	(0.44)	0.25	(0.43)	0.511	(0.019)
0.05	0.102	0.51	(0.54)	0.58	(0.56)	0.524	(0.022)	0.061	0.88	(0.46)	0.20	(0.41)	0.508	(0.017)
0.10	0.179	0.62	(0.65)	0.60	(0.61)	0.523	(0.022)	0.117	1.00	(0.61)	0.22	(0.43)	0.508	(0.017)
0.15	0.258	0.78	(0.81)	0.69	(0.70)	0.524	(0.022)	0.167	1.04	(0.70)	0.20	(0.42)	0.508	(0.016)
0.20	0.323	0.94	(1.09)	0.76	(0.81)	0.524	(0.022)	0.223	1.27	(0.92)	0.23	(0.46)	0.508	(0.016)
0.25	0.408	1.31	(1.40)	0.92	(0.95)	0.526	(0.022)	0.274	1.46	(1.21)	0.25	(0.49)	0.508	(0.016)
0.30	0.450	1.46	(1.76)	0.97	(1.07)	0.526	(0.023)	0.313	1.61	(1.58)	0.26	(0.50)	0.508	(0.016)
0.35	0.508	1.95	(2.32)	1.18	(1.27)	0.528	(0.023)	0.348	1.63	(1.81)	0.25	(0.51)	0.508	(0.015)
0.40	0.555	2.59	(3.17)	1.41	(1.53)	0.529	(0.023)	0.422	2.28	(2.35)	0.31	(0.58)	0.508	(0.015)
0.45	0.599	3.35	(4.49)	1.66	(1.87)	0.531	(0.024)	0.454	2.65	(3.11)	0.33	(0.61)	0.508	(0.015)
0.50	0.688	4.73	(5.75)	2.14	(2.17)	0.534	(0.024)	0.520	3.51	(3.92)	0.41	(0.71)	0.509	(0.014)
0.55	0.719	6.56	(8.20)	2.65	(2.72)	0.536	(0.024)	0.543	4.08	(5.12)	0.45	(0.77)	0.509	(0.014)
0.60	0.756	9.46	(12.01)	3.38	(3.43)	0.539	(0.025)	0.596	5.43	(6.73)	0.56	(0.89)	0.510	(0.014)
0.65	0.817	14.43	(18.38)	4.50	(4.43)	0.544	(0.026)	0.627	7.23	(9.44)	0.68	(1.03)	0.510	(0.014)
0.70	0.856	23.17	(29.54)	6.20	(5.83)	0.549	(0.026)	0.660	10.33	(14.44)	0.86	(1.26)	0.511	(0.014)
0.75	0.891	40.37	(51.18)	9.01	(7.96)	0.556	(0.027)	0.760	15.26	(21.52)	1.16	(1.57)	0.512	(0.013)
0.80	0.930	76.11	(93.50)	13.53	(11.09)	0.564	(0.027)	0.790	25.62	(37.09)	1.70	(2.22)	0.514	(0.014)
0.85	0.959	154.99	(164.53)	21.30	(14.83)	0.574	(0.025)	0.842	52.08	(79.22)	2.80	(3.42)	0.516	(0.013)
0.90	0.979	313.39	(259.70)	32.66	(17.57)	0.583	(0.022)	0.888	139.67	(222.82)	5.51	(6.28)	0.520	(0.013)
0.95	0.993	588.19	(327.90)	46.33	(16.55)	0.580	(0.019)	0.942	728.15	(1014.24)	16.18	(15.06)	0.528	(0.013)
0.99	0.998	795.26	(270.56)	54.52	(11.92)	0.570	(0.017)	0.989	3275.36	(2331.09)	41.07	(20.35)	0.532	(0.010)

Table E.18: AUC_***=0.8: Simulation results for the protected approach:** $m_e = 60$, $\Delta = 0.15$, $n = 100$ per group, $m = 1000$ and 6000.

	$m = 1000$							$m = 6000$						
FDR	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
0.00005	0.001	0.00	(0.00)	1.00	(0.00)	0.543	(0.000)	0.0003	0.67	(0.58)	0.33	(0.58)	0.514	(0.025)
0.0001	0.002	0.00	(0.00)	1.00	(0.00)	0.543	(0.000)	0.0005	0.60	(0.55)	0.40	(0.55)	0.517	(0.024)
0.0005	0.005	0.09	(0.29)	0.91	(0.29)	0.539	(0.013)	0.001	0.46	(0.52)	0.54	(0.52)	0.523	(0.022)
0.001	0.009	0.12	(0.33)	0.88	(0.33)	0.538	(0.014)	0.003	0.28	(0.45)	0.72	(0.45)	0.531	(0.020)
0.005	0.029	0.19	(0.40)	0.83	(0.40)	0.536	(0.017)	0.011	0.42	(0.50)	0.59	(0.49)	0.525	(0.021)
0.01	0.052	0.21	(0.41)	0.84	(0.44)	0.535	(0.017)	0.021	0.44	(0.50)	0.57	(0.50)	0.525	(0.021)
0.05	0.189	0.30	(0.51)	0.92	(0.63)	0.536	(0.019)	0.086	0.63	(0.60)	0.51	(0.53)	0.521	(0.021)
0.10	0.309	0.45	(0.70)	1.06	(0.86)	0.536	(0.021)	0.164	0.78	(0.69)	0.50	(0.57)	0.519	(0.021)
0.15	0.412	0.65	(0.94)	1.25	(1.08)	0.538	(0.022)	0.218	0.85	(0.84)	0.50	(0.60)	0.519	(0.021)
0.20	0.489	0.93	(1.32)	1.52	(1.39)	0.540	(0.024)	0.284	1.08	(1.01)	0.56	(0.66)	0.519	(0.021)
0.25	0.592	1.36	(1.72)	1.92	(1.66)	0.544	(0.025)	0.340	1.33	(1.31)	0.59	(0.72)	0.518	(0.020)
0.30	0.635	1.78	(2.37)	2.20	(2.08)	0.546	(0.026)	0.383	1.53	(1.66)	0.62	(0.77)	0.518	(0.020)
0.35	0.706	2.57	(3.20)	2.75	(2.47)	0.549	(0.027)	0.417	1.61	(1.94)	0.62	(0.81)	0.518	(0.020)
0.40	0.753	3.66	(4.34)	3.38	(2.96)	0.553	(0.028)	0.497	2.26	(2.51)	0.77	(0.93)	0.519	(0.020)
0.45	0.796	5.26	(6.19)	4.17	(3.62)	0.557	(0.029)	0.527	2.80	(3.43)	0.86	(1.05)	0.519	(0.020)
0.50	0.865	7.69	(8.55)	5.28	(4.26)	0.562	(0.029)	0.598	3.76	(4.28)	1.01	(1.18)	0.520	(0.020)
0.55	0.890	11.39	(12.29)	6.69	(5.18)	0.567	(0.029)	0.624	4.75	(6.11)	1.16	(1.39)	0.520	(0.020)
0.60	0.916	17.36	(18.04)	8.64	(6.30)	0.573	(0.030)	0.676	6.62	(8.45)	1.42	(1.61)	0.522	(0.019)
0.65	0.948	26.96	(27.15)	11.20	(7.57)	0.580	(0.029)	0.710	9.23	(12.35)	1.73	(1.96)	0.523	(0.019)
0.70	0.966	43.18	(41.53)	14.71	(9.06)	0.588	(0.028)	0.746	13.63	(18.25)	2.21	(2.43)	0.524	(0.019)
0.75	0.979	72.17	(67.15)	19.55	(10.78)	0.596	(0.025)	0.830	21.36	(28.37)	2.95	(3.02)	0.526	(0.018)
0.80	0.990	127.00	(113.16)	26.23	(12.69)	0.603	(0.023)	0.857	37.21	(50.01)	4.16	(4.06)	0.529	(0.018)
0.85	0.995	230.01	(179.49)	35.01	(13.79)	0.611	(0.018)	0.906	74.74	(97.52)	6.38	(5.72)	0.533	(0.017)
0.90	0.998	413.60	(260.73)	44.93	(13.24)	0.614	(0.016)	0.941	198.35	(258.15)	11.30	(9.09)	0.538	(0.016)
0.95	0.999	683.63	(290.36)	53.75	(10.03)	0.603	(0.018)	0.977	923.01	(1085.30)	25.27	(16.16)	0.546	(0.013)
0.99	1	851.06	(210.21)	57.90	(6.11)	0.592	(0.016)	0.997	3624.92	(2237.56)	47.69	(16.17)	0.545	(0.010)

Table E.19: AUC_*=**0.8: Simulation results for the protected approach:** $m_e = 60$, $\Delta = 0.15$, $n = 500$ per group, $m = 1000$ and 6000.

	$m = 1000$							$m = 6000$						
FDR	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)	$\hat{p}_s$	m_0^s	(SD)	m_e^s	(SD)	AUC	(SD)
0.00005	0.106	0.00	(0.03)	1.07	(0.29)	0.545	(0.005)	0.040	0.00	(0.00)	1.05	(0.23)	0.544	(0.004)
0.0001	0.157	0.00	(0.04)	1.12	(0.38)	0.545	(0.007)	0.057	0.00	(0.00)	1.07	(0.26)	0.544	(0.005)
0.0005	0.353	0.00	(0.07)	1.45	(0.79)	0.551	(0.012)	0.140	0.00	(0.05)	1.14	(0.43)	0.546	(0.007)
0.001	0.475	0.01	(0.07)	1.73	(1.06)	0.555	(0.015)	0.198	0.01	(0.08)	1.21	(0.54)	0.547	(0.009)
0.005	0.797	0.03	(0.16)	3.28	(2.21)	0.573	(0.025)	0.427	0.02	(0.14)	1.49	(0.94)	0.551	(0.014)
0.01	0.905	0.06	(0.25)	4.83	(3.00)	0.589	(0.029)	0.558	0.03	(0.18)	1.85	(1.32)	0.555	(0.018)
0.05	0.997	0.77	(0.99)	13.72	(5.22)	0.649	(0.030)	0.883	0.28	(0.57)	4.27	(2.76)	0.582	(0.027)
0.10	1	2.42	(1.94)	20.97	(5.67)	0.679	(0.024)	0.960	0.86	(1.15)	6.78	(3.85)	0.601	(0.031)
0.15	1	4.81	(2.98)	26.31	(5.74)	0.696	(0.020)	0.984	1.80	(1.86)	9.27	(4.51)	0.616	(0.031)
0.20	1	7.86	(4.09)	30.50	(5.63)	0.706	(0.018)	0.994	3.14	(2.73)	11.56	(5.06)	0.627	(0.030)
0.25	1	11.70	(5.42)	34.06	(5.47)	0.712	(0.016)	0.997	4.97	(3.78)	13.84	(5.33)	0.636	(0.027)
0.30	1	16.34	(6.94)	37.13	(5.27)	0.716	(0.014)	0.999	7.36	(5.01)	16.04	(5.60)	0.642	(0.026)
0.35	1	22.06	(8.70)	39.89	(5.03)	0.719	(0.013)	0.999	10.48	(6.51)	18.23	(5.80)	0.647	(0.024)
0.40	1	29.16	(10.95)	42.43	(4.78)	0.720	(0.012)	1	14.49	(8.40)	20.47	(5.90)	0.651	(0.022)
0.45	1	37.76	(13.65)	44.74	(4.54)	0.720	(0.011)	1	19.61	(10.78)	22.67	(6.01)	0.654	(0.020)
0.50	1	48.49	(17.21)	46.88	(4.30)	0.720	(0.011)	1	26.59	(13.82)	25.02	(6.12)	0.656	(0.019)
0.55	1	62.07	(21.75)	48.90	(4.04)	0.719	(0.010)	1	35.50	(17.71)	27.34	(6.17)	0.657	(0.018)
0.60	1	79.46	(28.11)	50.79	(3.78)	0.717	(0.010)	1	47.55	(23.22)	29.76	(6.20)	0.657	(0.016)
0.65	1	102.50	(37.12)	52.59	(3.46)	0.715	(0.010)	1	64.09	(30.34)	32.37	(6.19)	0.656	(0.016)
0.70	1	134.35	(51.07)	54.30	(3.12)	0.712	(0.010)	1	87.92	(41.18)	35.18	(6.18)	0.653	(0.015)
0.75	1	180.78	(74.66)	55.89	(2.76)	0.709	(0.009)	1	123.79	(58.25)	38.26	(6.09)	0.650	(0.014)
0.80	1	252.54	(112.20)	57.31	(2.30)	0.707	(0.009)	1	181.80	(87.66)	41.67	(5.97)	0.645	(0.014)
0.85	1	368.60	(166.75)	58.51	(1.77)	0.704	(0.008)	1	289.69	(148.98)	45.63	(5.76)	0.638	(0.013)
0.90	1	556.59	(232.71)	59.36	(1.18)	0.698	(0.014)	1	534.08	(324.82)	50.20	(5.31)	0.628	(0.013)
0.95	1	797.92	(216.91)	59.82	(0.64)	0.680	(0.018)	1	1421.33	(1058.18)	55.47	(4.30)	0.616	(0.012)
0.99	1	906.17	(125.61)	59.96	(0.29)	0.670	(0.012)	1	4134.20	(1967.25)	59.04	(2.18)	0.595	(0.016)

Bibliography

Abramovich, F., Benjamini, Y., Donoho, D. and Johnstone, I. (2006). Adapting to unknown sparsity by controlling the false discovery rate.
The Annals of Statistics, 34: 584–653.

Acion, L., Peterson, J., Temple, S. and Arndt, S. (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects.
Statistics in Medicine, 25: 591–602.

Anderson, T. (2003). An introduction to multivariate statistical analysis.
Wiley series in probability and statistics, third edition.

Bauer, P. (2008). Adaptive designs: looking for a needle in the haystack - a new challenge in medical research.
Statistics in Medicine, to appear.

Bauer, P., Pötscher, B. and Hackl, P. (1988). Model selection by multiple test procedures.
Statistics, 19: 39–44.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing.
J. R. Statist. Soc. B, 57: 289–300.

Benjamini, Y., Krieger, A. and Yekutieli, D. (2005). Adaptive linear step-up procedures that control the false discovery rate.
Technical Report.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency.
The Annals of Statistics, 29: 1165–1188.

Benjamini, Y. and Yekutieli, D. (2005). Quantitative trait loci analysis using the false discovery rate.
Genetics, 171: 783–790.

Brannath, W., Bauer, P. and Posch, M. (2002). Recursive combination tests.
Journal of the American Statistical Association 97: 236–244.

Bukszár, J. and Van den Oord, E. (2006). Optimization of two-stage genetic designs where data are combined using an accurate and efficient approximation for pearson's statistic.
Biometrics, 62: 1132–1137.

Dudoit, S., Shaffer, J. and Boldrick, J. (2003). Multiple hypothesis testing in microarray experiments.
Statistical Science, 18: 71–103.

Futschik, A. and Posch, M. (2005). On the optimum number of hypotheses to test when the number of observations is limited.
Statistica Sinica, 15: 841–855.

Genovese, C. and Wasserman, L. (2004). A stochastic approach to false discovery control.
The Annals of Statistics, 32: 1035–1061.

Goll, A. and Bauer, P. (2007). Two-stage designs applying methods differing in costs.
Bioinformatics, 23: 1519–1526.

Hommel, G. (1988). A stage-wise rejective multiple test procedure based on a modified bonferroni test.
Biometrika, 75: 383–386.

Jennison, C. and Turnbull, B. (2000). Group sequential methods with applications to clinical trials.
Chapman and Hall/CRC Press.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials.
Biometrics, 55: 1286–1290.

Li, L. and Hui, S. (2007). Step-wise variable selection and positive false discovery rate estimate in pharmacogenetics studies.
Journal of Biopharmaceutical Statistics, 17: 883–902.

Metnitz, P., Moreno, R., Almeida, E., Jordan, B., Bauer, P., Campos, R., Iapichino, G., Edbrooke, D., Capuzzo, M. and Le Gall, J. (2005a). SAPS 3 - From evaluation of the patient to evaluation of the intensive care unit. Part 1: Objectives, methods and cohort description.
Intensive Care Medicine, 31: 1336–1344.

Metnitz, P., Moreno, R., Almeida, E., Jordan, B., Bauer, P., Campos, R., Iapichino, G., Edbrooke, D., Capuzzo, M. and Le Gall, J. (2005b). SAPS 3 - From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission.
Intensive Care Medicine, 31: 1345–1355.

Miller, A. (2002). Subset selection in regression.
Chapmann and Hall/CRC, second edition.

Miller, R., Galecki, A. and Shmookler-Reis, R. (2001). Interpretation, design and analysis of gene array expression experiments.
Journal of Gerontology: Biological Sciences, 56: B52–B57.

Ntzani, E. and Ioannidis, J. (2003). Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment.
Lancet, 362: 1439–1444.

Ohashi, J. and Clark, A. (2005). Application of the stepwise focusing method to optimize the cost-effectiveness of genome-wide association studies with limited research budgets for genotyping and phenotyping.
Annals of Human Genetics, 69: 323–328.

Pencina, M., D'Agostino Sr., R., D'Agostino Jr., R. and Vasan, R. (2008). Evaluating the added predictive ability of a new marker: From area under the roc curve to reclassification and beyond.
Statistics in Medicine, 27: 157–172.

R (2005). R development core team: A language and environment for statistical computing.
R Foundation for Statistical Computing, Vienna, Austria.

Satagopan, J. and Elston, R. (2003). Optimal two-stage genotyping in population-based association studies.
Genetic Epidemiology, 25: 149–157.

Satagopan, J., Venkatraman, E. and Begg, C. (2004). Two-stage designs for gene-disease association studies with sample size constraints.
Biometrics, 60: 589–597.

Satagopan, J., Verbel, D., Venkatraman, E., Offit, K. and Begg, C. (2002). Two-stage designs for gene-disease association studies.
Annals of Human Genetics, 58: 163–170.

Shaffer, J. (1995). Multiple hypothesis testing.
Annual Review of Psychologie, 46: 561–584.

Shao, J. (1993). Linear model selection by cross-validation.
JASA, 88: 486–494.

Simes, R. (1986). An improved bonferroni procedure for multiple tests of significance.
Biometrika, 73: 751–754.

Skol, A., Skott, L., Abecasis, G. and Boehnke, M. (2006). Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies.
Nature Genetics, 38: 209–213.

Storey, J. (2002). A direct approach to false discovery rate.
J. R. Statist. Soc. B, 64: 479–498.

Storey, J. (2003). The positive false discovery rate: a bayesian interpreation and the q-value.
The Annals of Statistics, 31: 2013–2035.

Storey, J., Taylor, J. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach.
J. R. Statist. Soc. B, 66: 187–205.

Tsiatis, A., Rosner, G. and Metha, C. (1984). Exact confidence intervals following a group sequential test.
Biometrics, 40: 797–803.

Tusher, V., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response.
Proccedings of the national academy of sciences of the USA, 98: 10515–10515.

Van den Oord, E. and Sullivan, P. (2003). A framework for controlling false discovery rates and minimizing the amount of genotyping in the search for disease mutations.
Human Heredity, 56: 188–199.

Wang, H., Thomas, D., Pe'er, I. and Stram, D. (2006). Optimal two-stage genotyping designs for genome-wide association scans.
Genetic Epidemiology, 30: 356–368.

Zehetmayer, S. (2006). Two-stage designs for experiments with a large number of hypotheses. *Doctoral Thesis*.

Zehetmayer, S., Bauer, P. and Posch, M. (2005). Two-stage designs for experiments with a large number of hypotheses.
Bioinformatics, 21: 3771 – 3777.

Zehetmayer, S., Bauer, P. and Posch, M. (2008). Optimized multi-stage designs controlling the false discovery or the family wise error rate.
submitted.

Zehetmayer, S., Goll, A., Bauer, P. and Posch, M. (2007). Step by Step: mehr Effizienz mit neuen Studiendesigns.
Biospektrum 07: 754–755.